NBS
PUBLICATIONS

NBS SPECIAL PUBLICATION 616

U.S. DEPARTMENT OF COMMERCE / National Bureau of Standards

Validation and Assessment of Energy Models

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards[1] was established by an act of Congress on March 3, 1901. The Bureau's overall goal is to strengthen and advance the Nation's science and technology and facilitate their effective application for public benefit. To this end, the Bureau conducts research and provides: (1) a basis for the Nation's physical measurement system, (2) scientific and technological services for industry and government, (3) a technical basis for equity in trade, and (4) technical services to promote public safety. The Bureau's technical work is performed by the National Measurement Laboratory, the National Engineering Laboratory, and the Institute for Computer Sciences and Technology.

THE NATIONAL MEASUREMENT LABORATORY provides the national system of physical and chemical and materials measurement; coordinates the system with measurement systems of other nations and furnishes essential services leading to accurate and uniform physical and chemical measurement throughout the Nation's scientific community, industry, and commerce; conducts materials research leading to improved methods of measurement, standards, and data on the properties of materials needed by industry, commerce, educational institutions, and Government; provides advisory and research services to other Government agencies; develops, produces, and distributes Standard Reference Materials; and provides calibration services. The Laboratory consists of the following centers:

Absolute Physical Quantities[2] — Radiation Research — Thermodynamics and Molecular Science — Analytical Chemistry — Materials Science.

THE NATIONAL ENGINEERING LABORATORY provides technology and technical services to the public and private sectors to address national needs and to solve national problems; conducts research in engineering and applied science in support of these efforts; builds and maintains competence in the necessary disciplines required to carry out this research and technical service; develops engineering data and measurement capabilities; provides engineering measurement traceability services; develops test methods and proposes engineering standards and code changes; develops and proposes new engineering practices; and develops and improves mechanisms to transfer results of its research to the ultimate user. The Laboratory consists of the following centers:

Applied Mathematics — Electronics and Electrical Engineering[2] — Mechanical Engineering and Process Technology[2] — Building Technology — Fire Research — Consumer Product Technology — Field Methods.

THE INSTITUTE FOR COMPUTER SCIENCES AND TECHNOLOGY conducts research and provides scientific and technical services to aid Federal agencies in the selection, acquisition, application, and use of computer technology to improve effectiveness and economy in Government operations in accordance with Public Law 89-306 (40 U.S.C. 759), relevant Executive Orders, and other directives; carries out this mission by managing the Federal Information Processing Standards Program, developing Federal ADP standards guidelines, and managing Federal participation in ADP voluntary standardization activities; provides scientific and technological advisory services and assistance to Federal agencies; and provides the technical foundation for computer-related policies of the Federal Government. The Institute consists of the following centers:

Programming Science and Technology — Computer Systems Engineering.

[1]Headquarters and Laboratories at Gaithersburg, MD, unless otherwise noted; mailing address Washington, DC 20234.
[2]Some divisions within the center are located at Boulder, CO 80303.

Validation and Assessment of Energy Models

Proceedings of a Symposium
held at the National Bureau of Standards,
Gaithersburg, MD

May 19-21, 1980

Edited by:

Saul I. Gass

National Engineering Laboratory
National Bureau of Standards
Washington, DC 20234

Sponsored by:

Energy Information Administration
Department of Energy
Washington, DC 20461

and

Operations Research Division
Center for Applied Mathematics
National Engineering Laboratory
National Bureau of Standards
Washington, DC 20234

U.S. DEPARTMENT OF COMMERCE, Malcolm Baldrige, Secretary

NATIONAL BUREAU OF STANDARDS, Ernest Ambler, Director

Issued October 1981

Library of Congress Catalog Card Number: 81-600087

National Bureau of Standards Special Publication 616
Nat. Bur. Stand. (U.S.), Spec. Publ. 616, 248 pages (Oct. 1981)
CODEN: XNBSAV

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON: 1981

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402
Price $7.50
(Add 25 percent for other than U.S. mailing)

ABSTRACT

The Symposium on Validation and Assessment of Energy Models, held at the National Bureau of Standards (NBS), Gaithersburg, MD (May 19-21, 1980), was sponsored by the Energy Information Administration (EIA), of the Department of Energy (DOE), Washington, DC. The symposium was organized by NBS' Operations Research Division with a two-fold agenda: (1) to summarize the recent ideas and advances of model validation and assessment that have been applied to DOE energy models, and (2) to hold workshops on key open questions that are of concern to the validation and assessment research community. Speakers addressed current and future practices, the EIA model validation program, model structure and data, and model credibility. Full-day workshop sessions were held on the following topics: validating composite models, the measurement of model confidence, model structure and assessment, sensitivity and statistical analysis of models, and model assessment methodologies. This volume documents the symposium proceedings and includes the formal papers presented, discussant comments, panel discussions and questions and answers, and summaries of the issues and conclusions reached in the workshops.

Key words: Assessment; composite models; data quality; energy models; mathematical models; model confidence; model credibility; policy models; sensitivity analysis; validation.

CONTENTS

INTRODUCTORY REMARKS

Saul I. Gass
Symposium Chairperson

Good morning. I would like to welcome you to the Second Symposium on Validation and Assessment of Energy Models sponsored by the Department of Energy and the National Bureau of Standards.

The first symposium, which many of you were at, was held in January of 1979. That symposium emphasized the open issues and research aspects of energy model validation and assessment.

This symposium highlights the advances, new directions, and major research activities of the last year and a half. We have attempted to do this by limiting the invited paper sessions to a few specific topics and to spend almost two full days on research workshops.

I think I can best explain why we are here and why I think these symposiums are of great importance by quoting from the May 16, 1980 editorial that appeared in Science Magazine. It was written by Judge David L. Bazelon, Senior Circuit Judge of the U. S. Court of Appeals for the District of Columbia Circuit. The title of the editorial is "Science, Technology, and the Court." I quote:

> "I believe that the judicial responsibility is to monitor and scrutinize the administrative process. Our task is to ensure that the agency's decision-making is thorough and within bounds of reason. The agency's decisional record must disclose the evidence heard, policies considered and the agency's precise reasons for resolving conflicts in the evidence. This includes the basis for selecting one scientific point of view rather than another. This permits quality checks through peer review, legislative oversight, and public attention.
>
> "Only if decision-makers disclose assumptions, doubts, and moral and political tradeoffs can experts and citizens evaluate the administrative action. Only then can professional peer review bring to light new data or challenge faulty assumptions. Only then can Congress and the people gain sufficient understanding to permit meaningful debate of the valued choices implicit in regulatory action.
>
> "Acting independently about expert and political debate, courts can compel full ventilation of the issues on the record, as well as accustom decision-makers to the discipline of explaining their actions.

> "Finally, courts can ensure that all persons affected have opportunities to participate. The results should be an open process that can reveal gaps, stimulate research and thereby inspire more confidence in those affected, including the scientifically untutored."

I believe the work we are doing in validation and assessment of energy models, specifically, and models, in general, will go a long way in helping us achieve Judge Bazelon's goals.

This symposium is jointly sponsored by the Energy Information Administration of the Department of Energy and the Center for Applied Mathematics of the National Bureau of Standards. To offer a welcome from DOE, I would like to first introduce Dr. Lincoln E. Moses, Administrator of the Energy Information Administration of DOE. Then, on behalf of the National Bureau of Standards, Dr. Burton H. Colvin, Director of the Center for Applied Mathematics, will offer a few words of welcome.

WELCOMING REMARKS

Dr. Lincoln E. Moses
Administrator
Energy Information Administration
Department of Energy

My words are quite brief. I am here mainly as a listener. It is a pleasure to express the hopes for and the importance attached to this kind of work by the Energy Information Administration and to bid you welcome on behalf of one of the sponsoring organizations.

The conference itself gives good reason for hope because of the topics chosen and because of the participants who have been invited. Again, I bid you all welcome.

WELCOMING REMARKS

Dr. Burton H. Colvin
Director
Center for Applied Mathematics
National Bureau of Standards

On behalf of NBS and its Center for Applied Mathematics, I am happy to add my words of greeting and welcome to you.

This symposium is of special interest to NBS, generally, as all of our technical centers are rather deeply involved in collaborative work with all parts of the Department of Energy. In fact, these efforts constitute a relatively major and important fraction of NBS' other agency work. Also, interest in large scale mathematical modeling and methods to evaluate and assess such models has been a major thrust, in recent years, of the Applied Mathematics Center. We expect that to continue.

It is especially pleasant to have this collaboration between NBS and DOE. We share your thoughts that these are important problems, and our relationship with DOE has been pleasant and productive. It involves university organizations and a number of non-academic organizations, many of which are represented here. These are the kinds of collaborations which NBS has found extremely valuable.

I have a very special interest in adding my words of greeting. I wish you well in accomplishing the goals of this symposium.

ENERGY MODEL EVALUATION AND ANALYSIS:

CURRENT PRACTICE

by

David O. Wood*

1. Introduction

Formal efforts to improve the credibility and usefulness of energy policy models have proliferated since the first DOE/NBS Workshop on Validation and Assessment Issues of Energy Models.[1] That workshop was relatively unstructured, providing wide scope for presentation and discussion of both current activities and related issues, including the role of evaluation in policy modeling and research; the meaning of validation for policy models; and guidelines and standards for effective documentation of policy models and model-based studies. In contrast, the present Workshop emphasizes indepth collaboration of working groups meeting in parallel sessions. Such collaboration of workshop participants is obtained at the cost of reduced opportunities for project reports and open discussion by the full workshop. In partial compensation, Saul Gass has asked me to provide a brief survey of new activities and developments over the past year.

A convenient way to organize such a survey is to first mention new organizational initiatives and then to describe new projects and developments in agencies most prominently concerned with energy policy model development, application, and evaluation. These include the DOE Energy Information Administration (EIA) and the Electric Power Research Institute (EPRI). We conclude with a discussion of two significant activities in the past year: the EIA/DOE Workshop on the Measurement and Interpretation of Model Confidence and publication of the Texas National Energy Modeling Project (TNEMP) evaluation of the EIA Midterm Energy Market Model.

2. Organizational Initiatives

There have been several organizational developments in the past year relating to energy policy model evaluation and analysis including;

Energy Modeling Forum (EMF): The sponsorship of the EMF has been expanded to include the Gas Research Institute (GRI) and the Department of Energy in addition to EPRI, the founding sponsor.

Utility Modeling Forum (UMF): The EMF has spawned its first progeny in the EPRI-sponsored UMF, a development previewed by Cherry [1980] at the first Workshop. The UMF, administered by Booz-Allen, will employ the same approach to the organization and conduct of model comparison studies as the EMF, concentrating on such issues as load forecasting and

management, expansion planning and dispatch, and financial management. An early contribution of the UMF will be the publication of a Utility Model Notebook providing summary descriptions of utility based models.

Wharton Analysis Center: The University of Pennsylvania Wharton School has organized the Analysis Center for research in statistical and analytical methods relating to data and model development and evaluation. The Center, directed by Professor Larry Mayer, is presently concentrating upon energy models and associated data with sponsorship of the EIA Office of Validation Analysis.

M.I.T. Energy Model Analysis Program (EMAP): The M.I.T. Energy Laboratory has organized the EMAP for research in policy modeling and data analysis and to conduct policy model evaluations. The activities of EMAP are presently sponsored by EPRI, the Office of Analysis Oversight and Access (EIA) and the Office of Validation Analysis (EIA). In addition to conducting evaluation projects and related research, the EMAP will prepare the EPRI Energy Model Analysis Notebook as a source document for communicating results of evaluation projects, as well as model descriptions.

Texas Energy Policy Project (TEPP): The Texas Energy and Natural Resources Advisory Council (TENRAC) has organized the TEPP to support the council in various analytical and policy research activities. The objectives of the TEPP are summarized in Holloway [1980b, chapter 3]. Of particular interest is the objective of conducting an annual review and evaluation of the EIA's Annual Report to Congress and of national energy plans. The review and evaluation will be based in part upon use of the EIA Midterm Energy Forecasting Model, a current version of which is to be maintained by the TEPP. This model is presently used in supporting preparation of the Annual Report to Congress and of various energy information and policy studies requested of, or sponsored by, the EIA.

3. Energy Information Administration (EIA)

As has often been noted, much of the impetus for energy policy model evaluation derives from the Congressional concern as to the objectivity of the FEA's PIES applications. One product of this concern was the creation of the Professional Audit Review Team (PART).[2] The first PART report was extremely critical of the EIA's efforts to establish the credibility of their models. Thus,

> The credibility of OEIA's [now Energy Information Administration] models has not been established because documentation, verification, and validation have been neglected. Furthermore, publications describing the current models are scarce and procedures for public access to them are almost nonexistent. As a result, it is practically impossible for interested parties outside FEA to know whether OEIA's current models have been constructed properly and used correctly and thus whether OEIA's analytical products and forecasts can be used with confidence. (PART [1977, p. 31-32])

Partly in response to these concerns and findings, the EIA formed two offices, including the Office of Analysis Oversight and Access (OAOA)

under the Assistant Administrator for Applied Analysis, and the Office of Validation Analysis (OVA) under the Assistant Administrator for Energy Information Validation.[3] The functions and objectives of the two offices are complementary and serve the specific interests of the two Assistant Administrators. Thus, according to the 1978 Administrator's Report to Congress,

> The Office of Analysis Oversight and Access provides standards and procedures governing access, documentation, and technical review of applied analysis models and products. Primary attention is directed to evaluation of the data and methodology of analysis, and to the assessment and validation of models.
>
> The Office of Validation Analysis analyzes and evaluates the requirements for, and the consistency of, energy information systems and plans for specific validation studies. This office also is responsible for validations of projections and models.

3.1 Office of Analysis Oversight and Access (OAOA)

At the first Workshop, Lady [1980] described three broad classes of activities comprising the responsibilities of OAOA. These include documentation, evaluation projects, and improving public access to EIA models and associated data bases. OAOA has undertaken projects in each of these areas and, in addition, has sponsored a series of workshops to increase understanding of EIA models and modeling problems, and to improve credibility of EIA models and model applications.

Documentation Standards: The OAOA has formulated a system of standards for documentation of EIA-sponsored models and model applications. A preliminary set of standards was described in Lady [1980], and a revised standards are presently being developed. They are based upon OAOA's evaluation of the documentation needs of its various clients, supported by independent case studies.[4]

Archiving and Public Access: The OAOA has developed procedures for archiving the results of all analysis studies and associated models and data. These procedures are intended to ensure that any quantitative analysis conducted by the Office of Applied Analysis may be reproduced and verified either within EIA or independently. At present, archival files are maintained under the control of EIA. However, efforts are underway to develop an institutional capability at the Argonne National Energy Software Center.[5]

Evaluation Projects: OAOA has conducted or sponsored several evaluation projects. Projects recently completed include:

> An indepth evaluation of the Midrange Energy Forecasting System's econometric energy demand model (Freedman, et al. [1980]);
>
> An evaluation of the DOE Midterm Oil and Gas Supply Modeling System (NBS [1980]);[6]

An Evaluation of the Coal and Electric Utilities Model Documentation and Recommendations for Documentation Guidelines (Goldman, et al. [1979]);

Overview evaluation of the Short Term Integrated Forecasting System (Expert Panel, no report).

New and continuing evaluation projects of the OAOA include:

Indepth evaluation of the Short Term Integrated Forecasting System (NBS);

Review and evaluation of the PIES documentation (Professors Jerry Hausman and Carl Harris);

Comparative evaluation of the coal production cost functions for the National Coal Model and the Zimmerman Coal Supply Model (EMAP);

Comparative evaluation of the Baughman-Joskow Regionalized Electricity Model and the MEFS Electricity Submodel (EMAP).

In addition to evaluation projects, OAOA is conducting research on developing computer assisted methods for model analysis and evaluation. This research was previewed in Greenberg [1980] and provided the basis for a recent OAOA sponsored Workshop on Computer Assisted Analysis and Model simplification. Workshops were also organized to review the state-of-the-art of oil and gas exploration and production modeling, and to review and discuss issues in the measurement and interpretation of model confidence (discussed in section 5).

Analysis Review Procedures: The OAOA is responsible for organizing and conducting a review process for all analysis studies conducted by the Office of Applied Analysis. The process involves a combination of internal and external review and evaluation and is patterned on the peer review paradigm. The most ambitious review effort to date is the evaluation of the component models used in preparing the Annual Administrators Report for 1978, the subject of a separate session at this workshop.

Taken altogether, then, the OAOA has organized and implemented an extensive program of activities to evaluate and improve both the models and model-based studies of the Office of Applied Analysis. This program addresses all the issues raised in the original PART report [1977] -- including documentation, public access, and involvement of non-agency peers in evaluation -- and represents a serious committment to establishing the credibility of EIA models and model-based studies.

3.2 Office of Validation Analysis (OVA)

The Office of Validation Analysis (OVA) is organized under the Assistant Administrator for Energy Information Validation. Activities of OVA complement those of OAOA, emphasizing research on methods of data and

model evaluation. Current activities of OVA in energy information system evaluation include the following projects.

Princeton University Resource Estimation and Validation Project: A major effort at the Princeton University Department of Statistics and Geology over the past three years has been an indepth review of alternative methods for estimating the productivity of oil and gas exploration activity. This project has been of special importance because of EIA's legislated responsibility to develop independent estimates of oil and gas reserves. Approaches to survey design, data acquisition, and data validation must reflect an understanding of how such data may be used in developing informed projections. The objectives of the Princeton project include an analysis of the relation between the source data measurement system and the requirements and maintained hypotheses of alternative modeling methodologies.[7]

Wharton Analysis Center Evaluation of the Short-Term Integrated Forecasting System (STIFS): The underlying objectives of this recently initiated project are to develop and illustrate methods of data and model validation and evaluation using STIFS as a test model. STIFS is similar in concept to the Project Independence Evaluation System (PIES) in that demand relations are estimated statistically while supply and conversion activities are characterized by process submodels, all related via an integrating mechanism providing for the determination of equilibrium prices and quantities in energy markets. Initially, effort is concentrating on the demand submodels beginning with the demand for gasoline. Each model component is being examined in detail regarding such issues as a relation between underlying data measurement and the requirements of the model specification, and appropriateness of parameter estimation and calibration methods. Also, the consistency of the integrating mechanism will be evaluated in terms of the measurement system and accounting structure underlying the model. All these results will be combined and considered in developing formal measures of model confidence in relevant applications.

Oak Ridge National Laboratory Evaluation of the Long-Term Energy Analysis Program (LEAP): The objectives of this project are similar to those of the STIFS evaluation project, with LEAP serving as a test model to assist in "...development and demonstration of a methodology for evaluating energy-economic modeling codes and important results derived from these codes." (Alsmiller, et al. [1980], p.xi) While the objective is methodological, the focus differs somewhat from STIFS, with heavy emphasis on developing formal methods of sensitivity analysis for evaluating uncertainties in important data, and in measuring confidence limits of model forecasts. This latter problem is especially interesting since the methods of calibration used in LEAP -- engineering data and judgment -- do not provide a direct means of constructing probabilistic confidence limits for forecasts. To date an interim report has been prepared providing a detailed exposition of research objectives, and preliminary results for each of the main lines of research (Alsmiller, et al. 1980). This project represents perhaps the most ambitious and detailed research program yet conceived for developing methods of model analysis and evaluation.

M.I.T. Project in Model Evaluation and Data Validation: The objectives of this recently initiated project focus on validation techniques relating to the data/model interface. The emphasis of the research is to develop diagnostic techniques applicable to parameter estimation and system analysis, and to illustrate these methods on one or more EIA models. The diagnostic techniques for estimation are based in part upon Belsley, Kuh and Welsch [1980], extended to include estimation in simultaneous equation systems. System diagnostics methods will also be developed to measure relative importance of equations and specific parameters in model predictions, as well as summary measures of model forecasting performance. A third aspect of the research is source/model data transformations, and the implications for model specification and estimation in a particular model.

3.3 Summary

We began this section with reference to the congressional concerns about objectivity and "good practice" in the modeling and modeling applications of the EIA and its predecessor agencies, and to the findings presented in the first PART Report [1977]. The OAOA and OVA were in part created to address the congressional concerns and the combined programs represent a major effort. The success of their efforts may be partly judged from the second PART Report [1979] which finds things much improved regarding documentation, public access and evaluation. Thus,

> During 1978, EIA has taken various actions to improve the credibility of its energy models. PART believes that the development of a program plan for Applied Analysis and interim documentation standards are a major step in the right direction. EIA is also establishing controls over model changes, and it is sponsoring development of a model validation procedure, development efforts for model assessment which encompass validation, verification, and sensitivity testing procedures for models. It is also establishing procedures for public access to its models. (PART [1979, p. 40])[8]

4.0 Electric Power Research Institute (EPRI)

The EPRI Energy Analysis and Environment Division has conducted, supported, and encouraged a variety of projects and organizational initiatives concerned with increasing the understanding and use of energy policy models, and in developing models which are oriented toward particular problems and issues of EPRI's sponsors, the U.S. electric utility industry.[9] While EPRI has been somewhat less concerned than EIA with developing procedures for documentation standards and public access, they have made substantial progress in internal organization of the model development, evaluation, and application process, and in institutionalizing and supporting groups pursuing objectives of energy policy model analysis and application.

In a recent paper, Peck [1979] characterizes the program in terms of six "bridges" between research and modeling activities, and the clients of these efforts. The bridges include:[10]

Energy Modeling Forum,
Model Assessment Laboratory,
Heuristic Models,
Transfer Projects,
The Scenario Approach, and
Decision Analysis.

Energy Modeling Forum (EMF): Perhaps the best known of EPRI's initiatives to build bridges between modelers and model users is the EMF. Begun in 1976 with EPRI sponsorship,[11] the EMF has produced five major studies applying relevant models to specific policy issues, contributing understanding of the issue, improving user understanding of the models and increasing modeler sensitivity to the needs of policy analysts and decision makers.

The objectives, organization, and style of the EMF were described by Sweeney [1980a] at the first Workshop. Studies completed to data include:

Energy and the Economy (EMF 1),
Coal in Transition: 1980-2000 (EMF 2),
Electric Load Forecasting: Probing the Issues with Models (EMF 3),
Aggregate Elasticity of Energy Demand (EMF 4), and
U.S. Oil and Gas Supply (EMF 5).

Currently EMF 6, a study of world oil exploration and production models is underway, and a study of the macroeconomic effects of oil price shocks is being planned.

Energy Model Analysis Program (EMAP): In parallel with the EMF, EPRI has sponsored third party model evaluations by M.I.T. and others.[9] At an early EPRI workshop, the relation between EMF and third party analysis was described as follows:

> The panel described the role of third-party model analysis as a complement to Forum studies. The Forum must exploit the backroom concept of Forum operations, relying on the model developers to implement and translate the scenario specifications. The significant practical advantages of the procedure are achieved at the loss of the advantage of constructive independent investigation of model structure and operation. This activity supports the objectives of the Forum effort, but requires a different environment with intense involvement of individual analysts. The contributions of third-party assessment can be pursued independently.(EPRI [1977] p. 11-19).

Subsequently, EPRI sponsored the M.I.T. Energy Laboratory in third party assessments of the Baughman-Joskow Regionalized Electricity Model, the Wharton Annual Energy Model, and the ICF Coal and Electricity Utilities Model, in analyzing the process of organizing and conducting third party model analysis.

Recently EPRI's support of third party model analysis at M.I.T. has been expanded from a project to a program basis. The objectives of the

expanded activity include conducting evaluation studies and preparing an EPRI Model Assessment Notebook. Two projects are underway including an evaluation of the DFI Over/Under Capacity Planning Model, and of the Hirst Residential Energy Demand Model.

Transfer Projects: The organization, transfer, and application of knowledge in the form of research models and data is a difficult problem for EPRI managers. In 1978, EPRI asked a distinguished group of researchers to consider the problem and to recommend one or more approaches.[13] The group identified two approaches including :

> The iterative, problem-directed strategy, [in which] emphasis is on assembling information to meet a specific client need, using as necessary analytical tools tailored to the specific problem ; and
>
> the comprehensive modeling strategy..[which]..emphasizes linking together a broad spectrum of information in order to meet needs arising from a broad range of client problems. (Baughman, et al.[1978], p. 1-2).

Noting that in practice a mixture of both approaches would occur, the group recommended moving toward the iterative, problem-directed strategy.

Persuaded by the analysis underlying this recommendation and their own experience, EPRI Systems Program staff have developed the concept of the transfer project. In a transfer project, the analyst assembles information from existing research projects, may build a new model to integrate the information, and then transfers it to the client in a form tailored to the client's needs. The client provides feedback on important problems and information requirements which can then be used to guide the direction of future research and future transfer projects.

According to Peck, the ingredients for a successful transfer project are (a) to find a client with a problem that he is very interested in solving, (b) to include the client in a working group whose aim is the solution of that problem, and (c) to build a model which, although in its first phase may be somewhat crude, does nevertheless address the entire problem of the client. Peck [1979] cites as a mature instance of a successful transfer project the EPRI study, "Integrated Analysis of Load Shapes and Energy Storage." Another example would be the DFI Over/Under Capacity Planning Model. Although not originally conceived as a transfer project, the implementation of this model reflects important elements of the transfer project paradigm, especially planned interaction with model users in the model development phase. Several other projects are being planned as transfer projects, and experience with this approach is accumulating.

Success of the transfer project concept as a means of organizing applied modeling projects and studies will depend on several factors. Perhaps most importantly, transfer projects will require modelers and analysts with facility in moving between research and problem solving models. Models developed as part of a transfer project must be well

grounded in scientific knowledge and data, but because the emphasis is problem solving the resulting model may not be expressible in a scientifically refutable form. This suggests that an important aspect of successful transfer projects will be communicating to clients information about the confidence that should attach to model results.

A related concern is that the emphasis on transfer of knowledge and data to problem solving does not obscure the need to ensure the feedback from clients to modelers. The long run interests of clients will not be well served if modelers in transfer projects are primarily problem solvers as contrasted with agents of transfer and application of new knowledge and understanding. Of course it might be argued that this is preferable to a situation when knowledge is increased but no effective mechanism for the transfer exists. Clearly balance is required, and the success of the EPRI transfer project concept will depend upon the skill of its managers and clients in ensuring that this balance is achieved.

Modeling Approach and Study Organization: Under this heading I include the final three of Peck's bridges: heuristic modeling, the scenario approach, and decision analysis. By heuristic modeling Peck means the development and application of simple models which highlight an essential feature of some underlying process, perhaps characterized by a more detailed model. The simple model may be useful in providing insight into a key relationship as well as a basis for model comparison. Two prominent examples are the Hogan-Manne "elephant-rabbit" model of EMF 1, highlighting the relation between energy prices and economic growth, and Sweeney's simple model of oil exploration productivity in EMF.5, highlighting the essential relation between the shape of the oil finding curve, quantities discovered, and the price elasticity of cumulative discoveries.[14]

The System Group's own work (Peck [1979]) provides still another example of the uses of simple models. An important policy question in regulating sulphur dioxide emissions from new coal burning power plants is the cost of stricter regulation. EMF.2 considered this issue, among others, with the result that increased costs were not likely to be large,[15] although the variance among the models was significant. The Systems Group extended the results of the EMF.2, employing ICF Coal and Electric Utility model results in a graphical summary. The graphical model both summarizes detailed results, and highlights certain counter-intuitive results requiring further explanation.

Finally Peck identifies the scenario approach and formal decision analysis as important means of improving the effectiveness of model based policy studies. The scenario approach provides a structured method for the treatment of uncertain events and independent variables, either because of uncertainty in the realization for that variable (world price of oil in 1985), or because of controversy between study participants and/or clients. A decision tree framework provides both a means of structuring the process of scenario construction, and of organizing and presenting results as they bear on contentious points. Peck provides as illustration of this technique a study by Dale W. Jorgenson Associates [1979] of the economic consequences of a nuclear moratorium under alternative assumptions about several uncertain cost variables.

When the uncertainty in important variables can be objectively quantified in terms of probabilities, and when the study problem may be cast in terms of a particular decision, then decision analysis is a natural methodology to employ. As an important instance, Peck cites the efforts of Manne and Richels [1980] in conducting an economic analysis for the NASAP (Nonproliferation Alternatives Systems Assessment Program). One conclusion from the Manne-Richel's study is the difficulty of obtaining objective probability assessments from experts.[16] Of course, lack of confidence in the objectivity of probability estimates from experts who can "solve the problem backwards" does not affect the use of the decision tree framework for structuring studies, or the use of decision analysis as a means of identifying the critical uncertainties in a decision, regardless of their resolution.[17]

Summary: Cummulating experience suggests that EPRI's general approach to organizing problem oriented studies is successful, serving the needs of its constituency. How the EPRI approach evolves and the extent to which it is "transportable" to other organizations will be interesting to observe. Certainly groups with similar problems such as EIA, the Texas Energy Policy Project, GRI, EEI, and others would be well advised to study the EPRI process.

5. Measuring Model Confdence

During the past year, EIA and NBS have collaborated in sponsoring a workshop on measuring confidence in model predictions. The central importance of this issue to EIA is well summarized by the Administrator, Lincoln Moses, as follow:

> An energy model may have hundreds of equations and thousands of variables, constants, parameters. The formal structure of a forecast made from such a model is "If these thousands of details are correct, and if these equations correctly represent reality then the following consequent forecasts are correct." But some of those details have to be wrong, and some of the equations also. Then the forecasts have to be "wrong."... Why then offer them at all? The answer must be because we believe they are not seriously wrong. And suddenly their usefulness is seen to depend on assessing the likely degree of error. Thus, giving a useful idea of the uncertainty of a forecast is not a "refinement" -- it is central to the first order usefulness and meaning of the forecast. (Moses [1979], p.6).

A normative approach to summarizing information about the confidence to attach to a model has recently been proposed by Gass and Joel [1979] in an issue paper for consideration at an EIA/NBS Workshop on Establishing Model Confidence (October 4, 1979). Gass and Joel [GJ] propose four criteria and a scaling procedure to indicate the degree of confidence a client should attach to a model's predictions. The GJ criteria are documentation, verification, validation, and usability.[18] The scaling procedure is based on statements expressing

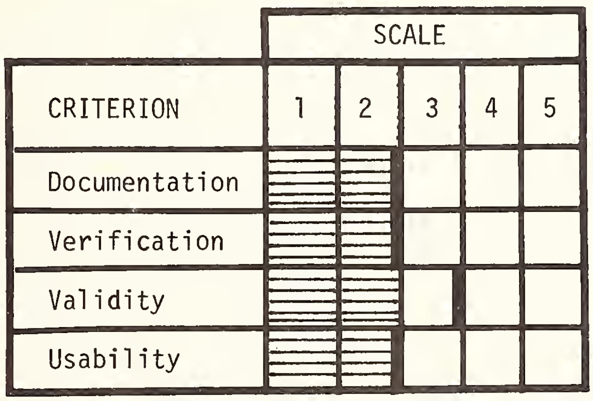
increasingly rigorous criteria conditions. Thus as an example of the most rigorous criteria for documentation they suggest the following:

> The documentation is sufficient for analysts, programmers, and non-technical personnel of another agency to obtain a detailed understanding of all aspects of the model. They should be able to run and interpret the model outputs, as well as make necessary modifications. (Gass and Joel [1979], p.18).

Such statements establish the basis against which the evaluation is to be conducted. If it has been agreed that such a statement represents the minimum acceptable criteria, then the findings of the evaluators are sufficient to determine if the minimum degree of confidence may be attached to the model. Thus, if there were five evaluative statments of increasing rigor attaching to the four criteria, then Figure 1 might be used to summarize an evaluation, where bold lines indicate minimum acceptable levels and darkened areas indicate attainment.

Figure 1

Summary Measure for Evaluation Of Model Confidence



Source: Gass and Joel [1979]

The formalism proposed by GJ is very suggestive, especially for summarizing evaluations of documentation and usability. However, I would see difficulties applying it to verification. If verification means, in part, matching the documentation to the actual computer code then verification tends to produce a binary result; documentation and code either match or they don't.

Employing the GJ formalism in summarizing efforts at model validation may also prove difficult, but is perhaps more tractable than for verification. The approach to implementing the formalism might proceed along the following lines. First the actual, intended and potential

applications of the model are described in sufficient detail to make clear the model accuracy required to provide sufficient discriminating power. Then the model is characterized in terms of structure, content, and information on predictive uncertainty. The requirements for accuracy, given the actual and intended applications, might be used to construct explicit and increasingly rigorous statements about model performance. The analysis of structure, content, and prediction uncertainty provide the source material for applying the criteria

Summary: The GJ proposal for summarizing informtion on model evaluation in a compact format, readily related to particular model applications, is very suggestive. We require experience with this approach to see how useful it will prove in practice, both to implement and to interpret.

6. Texas National Energy Modeling Project (TNEMP).

At the first workshop, Holloway [1980a] described the TNEMP, and presented a summary of the main results of that independent evaluation of the EIA Midrange Energy Forecasting System (MEFS). Since then, the final report has been published in two parts (Holloway [1979, 1980b]. Part II presents the detailed study results summarized by Holloway. Part I presents new material on:

further modeling and evaluation for Texas,
a report of the TNEMP National Advisory Board,
DOE review and comments, and
results of a TNEMP/EIA Workshop on substantive issues.

These materials represent an important contribution to the organization and practice of model evaluation. In a situation which at its beginning had a high probability of being:

acrimonious and adversarial,
a political rather than scientific exercise,
fragmented in terms of timing and location of published results and rebuttals, and
uninterpreted for those outside the process,

TNEMP, with EIA's cooperation, organized and implemented a process which:

maintained constructive tension between the EIA and TNEMP analysts,
maintained a scientific orientation,
ensured an outside evaluation of the process,
provided for rebuttal and discussion of controversial issues,
and published all results--evaluative, interpretive, and rebuttals--together.

There is no substitute for studying Part I of the TNEMP report (Holloway [1980b]) to appreciate how such a constructive outcome was achieved. In my view, however, there are five keys to the TNEMP success. First, the competence and scientific integrity of the

evaluation group was of high quality. Such a group would not lend their professional reputations to a non-scientific evaluation.

Second, considerable effort was devoted to establishing a constructive relationship between TNEMP and EIA, primarily via assignment of a technically competent, mature, and articulate EIA representative to provide liaison, and to prepare/coordinate comments and rebuttals to the findings of the evaluation group. This representative participated in all TNEMP review meetings. Thus the TNEMP process and intermediate results were well known to EIA, and so there were no surprises in the final report.

Third, the TNEMP organized an advisory group composed of persons knowledgeable about energy issues and modeling, with a majority of the memebers from organizations outside Texas. This group was responsible for advising the TNEMP director on matters of process; for preparing an interpretive report on the integrity of that process; and for preparing recommendations to the sponsoring organization, the Texas Energy and Natural Resources Advisory Council. In part this group served as a representative for those outside the process who were interested and concerned with the models and issues being evaluated and discussed.

Fourth, the TNEMP provided EIA the opportunity to prepare formal comments on the evaluation results, and published these comments in Part I of the report.

Finally TNEMP and EIA collaborated in sponsoring a workshop in which analysts in both groups made presentations and discussed the substantive issues of the evaluation, as well as more general research problems. A summary of the workshop presentations and discussion is also published in the TNEMP report.

All this attention to organizing and implementing a process which concentrated on both the form and substance of the evaluation has resulted in a study which is as self-contained and fair as possible. Regardless of ones views on the substance of the evaluation, the project serves as a model for organizing future independent model evaluation studies.

Summary: The TNEMP represents the most detailed and comprehensive independent model evaluation study to date. The organization of the project is worthy of careful study, providing a paradigm for future projects. Further experience with this effort is assured by the fact that independent evaluations of subsequent EIA Annual Administator Reports and modeling efforts are to be conducted as part of the on-going activities of the TEPP.

Notes

* Associate Director, Energy Laboratory, and Senior Lecturer in the Sloan School, Massachusetts Institute of Technology. This survey has benefitted from discussions with Saul Gass, James Gruhl, Douglas Hale, Kenneth Hoffman, George Lady, Martha Mason, and James Sweeney.

1. See Gass [1980] for the papers presented at that Workshop, and for transcripts of the discussions.

2. For a survey of congressional concerns and their impact upon energy legislation, see Mason [1979].

3. Dr. George Lady, Director of the Office of Analysis Oversight and Access, and Dr. Douglas Hale, Director of the Office of Validation Analysis, were both very helpful in providing materials and information relating to the activities of their respective offices.

4. Case studies include Gass [1979] and Goldman, et al.[1979].

5. Also a suggestion has been made to the EIA that it consider cooperating with the Texas Energy Policy Project (TEPP) in maintaining a current version of the modeling system underlying the EIA Administrators Annual Report to Congress, since TEPP is responsible for providing an evaluation of each such report. For discussion of this idea, see Holloway [1980, p. 65-66].

6. The NBS Energy Model Assessment project was organized "...to develop and apply standards and procedures for the assessment of analysis systems (models) utilized by the Energy Information Agency of the Department of Energy" (Gass, et al. [1980], p. 1). MOGSM was used as a test case for developing assessment methodologies for possible use by DOE and other modeling groups. The results of this effort are presented in ten reports submitted to EIA, and summarized in Gass, et al.[1980].

7. The results of the project are to be reported in nine technical reports and a summary final report. At this time three of these technical reports have been released including Mayer, et al.[1979a], Mayer, et al. [1979b], and Mayer, et al. [1980].

8. Italics in the original.

9. Earlier efforts are summarized in Wood [1979].

10. To this group we should add the Utility Modeling Forum.

11. As noted, sponsors now include DOE and GRI.

12. The REM and WAEM analyses are presented in Boshier, et al.[1979], with modeler comments in Chapter 4, and in Baughman [1980]. The perspective of EPRI is presented in Richels [1980]. An overview

of the CEUM is presented in Gruhl and Goldman [1980] with comments in Stauffer [1980].

13. The group included Martin L. Baughman, Edward Cazalet, Edward A. Hudson, Dale W. Jorgenson, David T. Kresge, Edwin Kuh, and D. Warner North.

14. In Sweeney's simple model, discoveries per unit of additional exploration are proportional to undiscovered reserves raised to some power, B. According to Sweeney [1980b], the value of B can not be closely bounded based upon current information, and so equally plausible values provide estimates of exploration productivity and price elasticity which bound the results from the more complex models.

15. The median cost increase for models participating in EMF.2 for a 9% SO_2 removal plus washing credit standard was 3% in 1985.

16. Manne and Richels find that probabilities relating to key events, such as consumer supply and demand growth, tend to be correlated with whether one is for or against plutonium fuel cycles.

17 See Cazalet [1980] for an elaboration of this point, and for a discussion of the application of decision analysis in model evaluation.

18. Gass and Joel [1980] later extended the criteria to include model definition, model structure, and model data. Documentation is now combined with usability. The essential features of their approach remain the same.

References

Alsmiller, R.G., et al.[1980], "Interim Report on Model Evaluation Methodology and the Evaluation of LEAP," (ORNL/TM-7245) Oak Ridge National Laboratory, Oak Ridge, Tennessee, April.

Baughman, M.L., E.G. Cazalett, E.A Hudson, D.W. Jorgenson, D.T. Kresge, E. Kuh, and D.W. North [1978], "Initiation of Integration," (EPRI EA-837) Electric Power Research Institute, July.

Baughman, M.L. [1980], "Reflections on the Model Assessment Process: A Modelers Perspective," in S.I. Gass [1980].

Bellesley, D., E. Kuh, and R. Welsch [1980], Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, John Wiley and Sons.

Boshier, J., J. Gruhl, R. Hartman, D. Kresge, E. Kuh, F. Schweppe, and D. Wood [1979], "Independent Assessment of Energy Policy Models," (EPRI EA-1071) Electric Power Research Institute, May.

Cazalet, E.G. [1980], "A Decision Analysts View of Model Assessment," in S.I. Gass [1980].

Cherry, B. H. [1980], "Electric Load Forecasting: Probing the Issues with Models," in S.I. Gass [1980].

Dale W. Jorgenson Associates [1979], "Industry and Economic Impacts of Restrictions on Generating Capacity," (Draft Report) Electric Power Research Institute.

Freedman, D., T. Rottemberg, and R. Sutch [1980], "The Demand for Energy in the Year 1990: An Assessment of the Regional Demand Forecasting Model," University of California, Berkeley, CA, May.

Gass, S.I. [1979], "Computer Model Documentation: A Review and An Approach," NBS Special Publication 500-39, U.S. Government Printing Office, Washington, D.C., February.

_________ (ed) [1980], Validation and Assessment Issues of Energy Models, Proceedings of a Workshop held at the National Bureau of Standards, Gaithersburg, Maryland, January 10-11, 1971. NBS Special Publication 569, February.

Gass, S.I. and L.S. Joel [1979], "Discussion Paper for the Workshop on the Measurement and Interpretation of Model Confidence," Presented at Workshop, October 4, 1979, NBS, Gaithersberg, Maryland.

_________________________ [1980], "Concepts of Model Confidence," (NBSIR 80-2053) National Bureau of Standards, June.

Gass, S.I., K.L. Hoffman, R.H.F. Jackson, L.S. Joel, and P.B. Saunders [1980], "The NBS Energy Model Assessment Project: Summary Overview," (NBSIR 80-2128) National Bureau of Standards, September.

Goldman, N.G., M.J. Mason, and D.O. Wood [1979] "An Evaluation of the Coal and Electric Utilities Model Documentation," Report to the EIA Office of Analysis Oversight and Access, MIT Energy Laboratory, September.

Goldman, N.L. and J. Gruhl [1980], "Assessing the ICF Coal and Electric Utilities Model," in S.I. Gass [1980].

Greenberg, H.J. [1980], "A New Approach to Analyze Information Contained in a Model," in S.I. Gass [1980]

Greenberger, M. [1980], "Humanizing Policy Analysis: Confronting the Paradox in Energy Policy Modeling," (Forthcoming in the Proceedings of this Workshop).

Holloway, M.L. (ed) [1979], Texas National Energy Modeling Project: An Experience in Large-Scale Model Transfer and Evaluation, Part II. Texas Energy and Natural Resources Advisory Council, Austin, Texas, August.

____________ [1980a], "The Texas National Energy Modeling Project: An Evaluation of EIA's Midrange Energy Forecasting System," in S.I. Gass [1980].

____________ (ed) [1980b], Texas National Energy Modeling Project: An Experience in Large-Scale Model Transfer and Evaluation, Part I. Academic Press, 1980.

Lady, G.M. [1980], "Model Assessment and Validation: Issues, Structures, and Energy Information Administration Program Goals," in S.I.Gass [1980].

Manne, A.S. and R.G. Richels [1980], "Probability Assessments and Decision Analysis of Alternative Nuclear Fuel Cycles," Energy Policy, March.

Mason, M.J. [1979], "Legislative Mandates for Energy Model Documentation and Access: A Historical Analysis," (MIT-EL 79-067WP), MIT Energy Laboratory, October.

Mayer, L.S., B. Silverman, S.L. Zeger, and A.G. Bruce [1979a], Assessing the Hubbert Approach to Modeling the Rate of Domestic Oil Production," Department of Statistics and Geology, Princeton University, September.

Mayer, L.S., K.L. Bell, E.S. Perwin, and P.C. Tittman, [1979b] "On the Use of Crude Oil Initial Production Statistics to Estimate Ultimate Production," Department of Statistics and Geology, Princeton University, October.

Mayer, L.S., R.A. Stine, B.W. Silverman, D.M. Snyder, S.L. Zeger, D.J. Venzon, and A.G. Bruce [1980], "The Use of Field Size Distributions in Resource Estimation," Department of Statistics and Geology, Princeton University, January.

Moses, L. E. [1979], "One Statistician's Observations Concerning Energy Modeling," Keynote Address to NATO Conference on Energy Modeling, Brookhaven National Laboratory.

Peck, S.C. [1979], "Six Bridges Between the Builders and Users of Energy Models," IGT Symposium on Energy Modeling, Colorado Springs, Colorado, August.

Professional Audit Review Team [1977], Report to the President and the Congress: Activities of the Office of Energy Information and Analysis, U.S. General Accounting Office, Washington D.C., December.

______ [1979], Report to the President and the Congress: Activities of the Energy Information Administration, U.S. General Accounting Office, Washington, D.C., May.

Richels, R. [1980], "Third Party Model Assessment: A Sponsors Perspective," in S.I. Gass [1980].

Stauffer, C.H., Jr [1980], "Developing, Improving, and Assessing the ICF Coal and Electric Utilities Model," in S.I. Gass [1980].

Sweeney, J.L. [1980a], "Energy Model Comparison, An Overview," Prepared for Symposium on the Outlook for Large-Scale Energy Models, AAAS Annual Meetings, San Fransisco, CA, January 4-5.

______ [1980b], "The Energy Modeling Forum: An Overview," in S.I. Gass [1980].

Wood, D.O. [1980], "Model Assessment and the Policy Research Process: Current Practice and Future Promise," in S.I. Gass [1980].

HUMANIZING POLICY ANALYSIS: CONFRONTING THE PARADOX IN ENERGY POLICY MODELING

Martin Greenberger
Johns Hopkins University
Baltimore, Maryland

SUMMARY

Energy policy models (and energy policy studies generally) are much in evidence in Washington and State capitals these days. Their presence grows with each advancing year. The needs of policymakers for analytical assistance in the face of complex, multifaceted policy problems have never been greater. Yet policy analysis is still young in its development and has a difficult time living up to the expectations and requirements of its users. Like mathematically precocious youngsters, policy models--the most structured form of policy analysis--are not well understood by even sophisticated nontechnicians and a mystique surrounds them. Many of those with most to gain from policy models--and with the real-world wisdom needed to bring them to maturation and practical usefulness--view these analytical prodigies with discomfort and even disdain. The high degree of currency which policy models enjoy is matched by the considerable skepticism (and sometimes contempt) with which they are regarded.

We examine the reasons for this apparent paradox and consider how to deal with it. Several organized activities have recently begun to appear aimed at dissolving the mystique surrounding policy models and making them generally more intelligible and useful. After reviewing these activities and describing the forms they take, we ask whether their benefits are being perceived and appreciated by policymakers. The efforts are clearly of educational value to the technical participants. We explore what can be done to ensure that their value is extended to decisionmakers as well, and to others whose interests would be served by having policy models play a more natural and easily assimilated role in human affairs.

INTRODUCTION

> "When the history of economic policymaking in the turbulent late 1970's is written, an important part of the story will be about the ever-widening impact of econometric models on Federal policy decisions. The wondrous computerized models--and the economists who run them--have become new rages on the Washington scene."[1]

But what are "rages" one year can become passé and be discarded the next. Policy models are in their adolescence, their popularity is tempered by doubt and questioning, and their footing is not secure. What is to become of them in the future?

The fact is that policy models, despite their "ever-widening impact," are not yet accepted and understood in realistic terms by policymakers--their ultimate users. Whether this is an innate and inevitable characteristic, or a deficiency that may be corrected by further development, can better be decided after a review of the nature of policy models and the images they convey.

MODELING IN PERSPECTIVE

Modeling for policy decisions had its origins in the 1930's.[2] As practiced at the present time, policy modeling is a confluence of several separate branches of methodological development, of which econometric modeling is one. Each branch has had its own rich and interesting history.[2] Policy modeling borrows from modeling in the sciences--using some of the same ideas and mathematical paraphernalia. But the differences between these two types of creative activities are more significant than the similarities.

The purpose of modeling in the sciences is typically to express the theories and understanding of the modeling investigator in a well-phrased form (often mathematical, but not necessarily so) that facilitates <u>testing</u>, <u>exploration</u>, and <u>communication</u> in order to develop or reinforce the conceptual structure of a discipline. Modeling in a policy context has a different purpose. It is oriented toward <u>decisions</u> rather than disciplines. Policy modeling is fundamentally <u>pragmatic</u>, not theoretical. Its primary goal is <u>not</u> building up the structure of knowledge in a field of learning; its overriding objective is forming a basis for policymaking.

The specific intent of policy modeling is to provide an instrument for comparing the effects of alternative policy actions. Policy modeling is designed in a general sense to lay the groundwork for the intelligent and informed making of decisions. But to achieve this aim, policy modeling must lend itself to testing and exploration by others than its developers. It must be possible to communicate the rationale of policy models as well as the results. Policy modeling shares these obligations with modeling in the sciences. Both forms of modeling ultimately should be outward-facing and intuitively understandable to nonmodelers.

If a policy model cannot be tested, explored, and comprehended by persons not part of its development, one might expect its future to be brief and its use restricted. Yet, in their adolescence, policy models have often been objects of blind reverence and admiration or equally blind awe and mistrust. They have been accepted or rejected because of the personal qualities and standing of the person presenting the results--and because of the predisposition of the person receiving the results--more than because of characteristics of the models themselves. And their role is expanding.

> "Just this year some costly farm bills were strangled in their cradles after the econometric models pronounced them expensive and inflationary.... Nearly every section of the $18.7-billion tax cut passed by Congress last month was tested in models for its effect on revenues and economic growth. The controversial reduction in capital-gains taxes was buoyed up against strong White House opposition by computer printouts showing that it would have a positive effect on investment, stock prices, and even Treasury receipts."[1]

With maturity comes the call for greater accountability. Policymakers and their staffs are starting to ask why models produce the results they do.

> "The first phase was an open-faced deference to the precise multi-digited descriptions of the future that the models poured forth; the second phase was anger when favorite oxen got gored by the sharp numbers from the computers.... In the third phase, the legislators and bureaucrats have begun questioning the assumptions and even the ideological biases built into the models."[1]

In the energy field, policy problems are diverse and intertwined, with aspects of economics, technology, resources, and regulation all having a bearing. Energy is a ripe field for modeling, and many energy policy models have been actively developed and used since the days of the Mideast oil embargo.

> "On more than one occasion, the models have contradicted pronouncements by senior government figures. A small model at the Council of Economic Advisers, for instance, indicated that . . . the Secretary of Energy was talking through his hat . . . when he predicted that disastrous consequences would flow from last winter's coal strike . . . The model, it turned out, was correct."[1]

What was "correct," of course, was not the model, but a particular conclusion drawn from the model. Models may be clever, illuminating, perceptive, and useful, but they are never "correct" in any absolute sense. As deliberate simplifications and reductions of reality, they must always fall short in their representations of the systems they serve as proxies.

COUNTER MODELING

Policy models are products of the assumptions that go into them. As with any source of policy advice, they should be checked and challenged. Lyndon Johnson was said to probe the reasoning of his policy advisors by asking them to present the most persuasive argument they could supporting the opposite side of the position they espoused.[3] Policy models are well suited for this exercise. The practice of altering the assumptions of a model--or using a different model--to challenge the conclusions reached in a modeling study is known as "counter modeling."[2] More than a logical exercise, it is becoming a serious tactic in courtroom cases and policy debates.

To cite one example, the same model used in a policy study of nuclear energy to question the necessity for the breeder reactor program and reprocessing was subsequently used with different assumptions to show the high risk to the economy to be incurred by suspension of breeder development.[4] What became clear from the exchange was the extent to which the conclusions depended upon assumptions about the availability of uranium, the future demand for electricity, and the use of coal.

Some decisionmakers are not so eager to have both sides of an argument laid out before them. Just as they may prefer the "one-handed" lawyer who does not invariably interrupt a line of reasoning with "on the other hand," so they look for the analysis that leads unequivocably to the conclusions with which they feel most comfortable. Tracing the results of alternative assumptions makes them impatient or confused. Similarly, a judge trying a case into which counter modeling is introduced may have trouble understanding how one model can produce two very different outcomes. A congressional committee holding hearings on proposed new energy legislation could have conflicting stories of a half dozen models read before it. The Congressmen may not be disposed to believe any of the models. In the political climate of Capitol Hill, disagreeing modelers seem to be "grinding their axes" like everyone else.

These realities appear paradoxical. There is need for policy analysis and a desire to reach for its most sophisticated and highly developed form--policy modeling. But policy models are "many-handed" tools that can be adopted, adapted, or rejected. They are useful in making a case, supporting an argument, or defending a position. Yet their very versatility sows the seeds not only of their propagation but of possible skepticism as well. Policy models are versatile but explicit, and therefore very vulnerable. They become subjects of attack and misunderstanding. Their intrinsic dependence on assumptions--and therefore on the "spirit of the times"--causes them to be mistrusted, dismissed, or impugned.[5] At the same time that modeling studies are growing in number the uncertainties surrounding them are increasing apace.

CREDIBILITY

In a paraphrase of a modern commercial, it may seem nowadays policymakers are "using model analysis more and believing it less." In some ways, this is true. But it is not the complete picture. Credibility is a problem, but not an insurmountable one. When an analysis is viewed in appropriate perspective and clearly centered on the key elements of a question or disagreement, it can be helpful to both sides of a dispute in clarifying the issues between them and providing a framework for discussion. This is illustrated by the following vignette, based on a current matter of great interest in electric utility capacity planning.[6]

> Focusing the Argument. Members of a regional pool of electric utilities have been at odds with their local public utility commissions about the amount of expansion needed in generating capacity to meet the anticipated power needs of the region. Two nuclear base-load plants have been canceled and there has been a three to four year delay in construction of 12 other plants. The argument has taken the form of *ad hominem* attacks

on opponents, with accusations of greed and obstructionism clouding the issues. Recently, a model was developed under the aegis of the Electric Power Research Institute (EPRI) by a private consulting firm, Decision Focus, Incorporated, in an attempt to structure and quantify the dispute. Both the individual utilities and the public utility commission are using the model, and it is helping them to clarify their points of difference. The model's two key assumptions have to do with the expected load in the 1980's and the implicit cost of not being able to supply electricity to customers. The model provides a means for calculating the probability of service interruptions at different reserve margins under uncertain loads, and assigns a consumer cost to this probability. Despite a recent fall-off in electricity demand, the utility members predict a serious shortage in generating capacity during the 1980's. Federal energy officials basically agree with this outlook. The utilities maintain that it is much more costly for them to be caught short in generating capacity than to be oversupplied. Both sides are using the model to explore their assumptions about future load growth and the impact of under-capacity. They are now investigating the validity of the model and its response to various parameter settings. The argument is taking on a more reasoned tone and there is the beginning of cooperative planning.

Crucial to the success of the capacity planning model is its credibility and acceptance by the users. An analysis that has had more difficulty gaining acceptance is the reactor safety study produced in 1975 for the Atomic Energy Commission. The results of that study, cited often by proponents of nuclear energy to reassure the public about the safety of nuclear reactors, were drawn into question four years after its release by the Nuclear Regulatory Commission (NRC) for being overly confident on its low estimates of the probability of reactor failure.[7] Coming just months before the Three Mile Island incident, the NRC action may seem prophetic and timely. But Harold Lewis, the scientist who led the review group critique that was the basis for the NRC action, faulted the NRC for not using the original study appropriately.

> "The NRC should have jumped at the opportunity of knowing for the first time what most accident sequences were, what the relative importance was, what they ought to be doing with inspection time, regulatory time, rule-making, research . . . "[8]

In fact, the original study had pointed out that a high-risk source of trouble with reactors were small-break, loss-of-coolant, and transient-initiated accidents, the kind of minor problems that can lead to a major problem of the variety that occurred at Three Mile Island. Yet the study was never significantly used as a diagnostic tool. Nuclear critics claim that it was used instead as a tranquilizer, and they charge that was its intent from the start. It is not necessary to be this cynical to conclude that a valuable opportunity was missed in not applying the study to identify problem areas. The consequences were costly. To its credit, and consistent with the findings of both the Lewis review group and the Kemeny Commission that investigated the causes of the Three Mile Island incident,[9] the NRC began conducting a program to apply the probability risk assessment methods of the original Nuclear Reactor Safety study to identify "risk outliers" or weak points in the design and operation of reactor systems.

SCAPEGOATING

Sometimes a policy analysis is given more credit for producing an attitude or action than is warranted, especially when the subject of study is unpopular or controversial. A policy model presents an easy, impersonal target, like the computer blamed for a breakdown in customer service. "The computer must have fouled up." The following vignette presents an example of this modern-day form of scapegoating drawn from the early days of U. S. reaction to the "energy crisis" invoked by the Arab oil embargo of 1973-74.[10]

> A Charge of Bias. The Energy Research and Development Administration (ERDA), set up to fund programs to reduce the nation's dependence on foreign oil, issued its first planning document urging major support for advanced technologies. The report, hardly mentioning conservation, is attacked vociferously by those who consider conservation as a first order of business in dealing with the energy problem. They charge that the report caters to the interests of the energy industries and the nuclear establishment in promoting high technology options to the exclusion of conservation. ERDA responds that it was under great time pressure in producing its plan, and the neglect of conservation was inadvertent.
>
> ERDA used an energy system model developed at the Brookhaven National Laboratory as an aid in drafting the plan. One outspoken critic claims that the model was misused to buttress ERDA's position. The analysis assumed a fixed set of end-use demands for such energy services as the number of dwellings to be heated, passenger miles of travel, and square feet of commercial floor space. End-use efficiencies, such as for home furnaces and building shells, were increased in a conservation scenario that was run to reflect energy reductions from possible improvements. Kenneth Hoffman, the model's developer, notes that this conservation scenario had the largest impact of any of the scenarios used in the analysis by a wide margin. The conservation runs were reported in an appendix to the report. Hoffman comments on ERDA's downgrading of these runs. "It is quite common in policymaking for other important considerations to outweigh those treated in a quantitative analysis. This does not necessarily indicate a misuse or deliberate manipulation of analysis." Hoffman believes the government's position is that R & D in the conservation (end-use) area is a private sector responsibility and does not require new government initiatives.[11]
>
> After a lively exchange, Hoffman and the critic agree on the desirability of holding a workshop to air the issues. One of the important questions is the extent to which the model, by its nature and content, itself influenced the conclusions and coverage of the report.

The workshop agreed to by the two never took place as it turned out, but a meeting similar to what they had in mind did convene shortly before to consider another modeling study. The background for this meeting is outlined in the following vignette. It is an illustration of the bewilderment that can result from the subtleties and versatility of a model.[12]

The Effect of Energy on the Economy. The Hudson-Jorgenson model was the first to combine energy price substitution effects with a representation of the national economy. It has been used in a wide variety of applications. The model can portray, for example, higher energy prices leading to reduced energy consumption through a shift to energy-conserving technologies. Some observers wonder how it is that the model's results seem to change from one application to the next. The Ford Energy Policy Project cited runs of the model to show that a major reduction in energy consumption was possible without seriously injuring the growth prospects of the national economy or increasing unemployment.[13] Yet in an industry-sponsored study, the model displayed substantial energy cutbacks accompanying a fall in national productivity and a retreat in economic growth. Jorgenson argues that there is no inconsistency. The results are entirely reasonable, he explains, in terms of the different ways the model was applied and the different variables held constant in the two sets of runs. A meeting is called by EPRI to examine the applications in detail and to obtain a better understanding of the model. Some of those invited get the mistaken impression that the meeting is to be a "kangaroo court" and decline to participate. Many others come prepared to support the model and its uses.

MODEL ASSESSMENT

The EPRI meeting, held in early 1976, was one of the first attempts to bring experts together to probe the workings and examine the uses of an energy model. The idea was new and a less than diplomatic approach was taken in assembling the group and organizing the meeting. It led to misunderstanding, irritation, and a loyal defense of the model and its creators. But the notion of model assessment was taking hold, and awareness of the need for this kind of activity was growing.

A much more extensive attempt at model assessment occurred two years later in the State of Texas. The lieutenant governor was unhappy with Federal projections of the supply response likely from the deregulation of natural gas prices and the emphasis given to conservation over production incentives in President Carter's National Energy Plan (NEP). The large and complex Department of Energy (DOE) model (PIES) that had produced the projections justifying NEP was made the subject of an assessment designed to scrutinize the properties of the model and expose its limitations.[14] A second purpose of the assessment was to provide information and possibly counter models for the State's own use in the adversarial debate over national energy policy. The debate goes back 25 years. At issue were the appropriateness of Federal wellhead price ceilings for natural gas sold in interstate commerce and the ability of state utilities to continue using state-produced natural gas to generate electricity. The stakes were high and so were the feelings.

The Texas assessment was well organized, with researchers from several Texas universities working under the aegis of a State office and a national advisory board. In preparation for the evaluation, the PIES model was successfully "transferred" to a computer center at Texas A&M University for operation there. A manuscript reviewing the model was prepared for publication.[14a] One

outcome of the assessment was a strong recommendation for continuing review of DOE modeling work in the public interest.[15] The desirability and feasibility of model transfer was the subject of a symposium organized at an annual meeting of the American Association for the Advancement of Science.[16]

THIRD-PARTY MODEL ANALYSIS

Model assessments are only one of the means adopted in attempting to address questions such as those posed in the preceding vignettes. These questions can to a certain extent be answered by the developers and users of the models themselves. But developers and users cannot be entirely objective and discriminating critics. The developers want to get their models working and accepted. They are understandably admirers of their creations and are not always in the best position to perceive the shortcomings of what they have wrought.

Similarly, users of a model may feel a special allegiance or commitment to it, particularly if they have borne the costs of its development. They will be influenced by their association with the developers and by any support the model lends to their policy positions.

The vignettes portray models being used in policy debates and being perceived by some of the parties in these debates as a threat or challenge. The injured or perplexed parties want to have a closer look at the models and find out more about how they came to produce such alien and puzzling results. The models become targets for investigation. "Third-party" model analysts, aligned organizationally neither with the model builder nor model user, are in a good position to undertake these investigations.

Model analyses by "third-party" model analysts (sometimes working together with model developers and users) fall into two categories: 1) model assessments focusing primarily on the model, and 2) forum analyses focusing more on policy issues and uses of the model.[17] An example of model assessment is the work of the Model Assessment Group at the Massachusetts Institute of Technology under an EPRI-funded project.[18] An example of forum analysis are the studies of the Energy Modeling Forum (EMF), set up by EPRI and Stanford University in 1976, with funding now coming from several sources.[19]

In a typical model assessment, a single model such as PIES is examined in its many ramifications. A series of computer runs is made to explore and evaluate the model's capabilities, limitations, adequacy, and realism. One model is run over several issues, in a manner of speaking. The model may contain numerous components or sub-models with different characteristics.

In the forum analysis, on the other hand, a number of models are applied to one set of questions relating to a single subject of inquiry. That is, many models are run on a single issue. The first EMF study on "Energy and the Economy," for example, explored the effects of reduced energy consumption on the performance of the national economy by running the Hudson-Jorgenson model and four other models under a set of common assumptions agreed to by the modelers.[20] Later EMF studies took up successively the issues of: coal in transition,[21] electric load forecasting,[22] energy demand elasticities,[23] oil and gas supplies,[24] and the price of world oil.[25]

Model assessment examines a model over the full range of its capability. Forum analysis, in contrast, concentrates on a focused set of questions, applies a set of models to these questions, compares the results, and probes the differences. Model assessment is designed specifically to evaluate a model and understand it. Users are not a necessary part of the operation, although they may prove helpful. In forum analysis, they are essential. The aim of forum analysis is to broaden user insights and understanding, not only of the models, but of the issues as well. As a by-product, modelers who participate along with the users in the forum process gain a fuller awareness of what their models can and cannot do and what the user does and does not need. The forum provides a comparative commentary on the models it employs. But evaluation of these models is not its primary purpose.

Both kinds of analysis focus on specific policy issues. It makes little sense to analyze a model in the abstract, divorced from the concrete uses for which it is intended or could be applied. The two forms of analysis are complementary. Their differences are a matter of degree and style. Each has its own set of objectives and each contributes in its own way to answering questions about the models as well as the issues to which the models are applied.

Awareness of the need for third-party model analysis is significant and growing. There is by now a large number of model analyzing activities underway. Model analysis appears to be taking root. It could become a recognized, valued, and permanent element in policy research over the long term, not only in the energy field, but in other policy areas as well.

Whether an institutionalization of model analysis will come about is an interesting question. There are indications that the development is heading that way. Experience with model analysis to date has served to reinforce the conviction that the role it plays will be "central . . . in efforts to make policy models more useful."[2]

EMERGING FROM ADOLESCENCE

The adolescence of energy policy modeling may be coming to an end with the call for greater accountability reflected in the growth of third-party model analysis. But more is needed before policy modeling can reach healthy maturity. The seeming contradiction between increasing popularity and increasing skepticism points up the fact that policy modeling has yet to become well-adjusted and at ease with its complementary culture, policymaking.

It is naive to expect policymakers to speak the specialized language and adopt the technique-oriented criteria of policy modelers. These two cultures are different--in background, values, and operating styles. Policy analysis is disciplined, exacting, and tends to be highly stylized. Policymkaing is outward-looking, intuitive, and can be myopic. But both activities are integrative and problem-directed. They should be mutually supporting, rather than at odds or competitive. They can learn from each other and can take advantage of one another's insights and expertise.

Many countries of the world do not experience the tension and communication problem between policy analysts and policymakers found in the United States.[26] The reason is that policy analysts in other countries often work directly for policymakers who initiate the work and establish its parameters. This is not to say there is not direct reporting in the United States as well. But much of the most influential policy analysis done in this Nation of highly fragmented, multifarious interest groups takes place in universities, research organizations, and "outside studies" removed from the point of decision and without organizational ties to the decisionmakers. Even policy research arms of the Federal Government are often called upon to serve a multiplicity of masters inside and outside of government.

Understanding between policymakers and policy analysts is aided by the construction of tunnels and bridges between them, e.g., the Energy Modeling Forum and the capacity planning study discussed earlier.[27] It is as though there were a dense thicket or a hard mountain through which both policymakers and analysts must cut to find their way. The two groups begin penetrating on opposite sides. If they meet in the middle, they have one tunnel that leads them through. If they do not meet, they have either no tunnel or two separate tunnels likely to take them in divergent directions and accentuate the gap between them.

Why must the two groups work from opposite sides? The two cultures are unalike in many ways. But there are people with the knowledge and temperament to be able to operate effectively in both cultures: policymakers with strong analytical bent and solid technical training, and policy analysts who have carried executive responsibility and have a special sensitivity for the concerns and constraints of the decisionmaker. In addition, there are many people between these polar stereotypes who are able and well-equipped to perform an intermediating role.

There are enough such people to make it possible to launch a cooperative effort starting from the same side, drawing on expertise as needed from both cultures. From which side should such an effort begin?

The Energy Modeling Forum was an approach to the tunneling project from the modeling side. It has produced substantial benefits for its technically oriented participants. But it has had only partial success in engaging the interest and participation of nontechnically oriented policymakers. This is not to say that policymakers have not been involved and influential in EMF activities. They have. A senior advisory panel of top-level policymakers oversees the Forum and has been a strong force in guiding and proposing EMF studies. The insights and wisdom of members of this panel have had a decisive impact on more than one occasion. It was their critique, for example, of the preliminary results of the first EMF study on "Energy and the Economy" that focused attention on the effects of energy consumption on capital formation. This concern turned out to have major significance in the later findings.[20]

But the senior advisory panel, by design, functions separately from the working groups who conduct the EMF studies, and it meets much less frequently. The advisors are sounding boards and idea generators more than participants. When nontechnical people do attend the meetings of the working groups, they are often overwhelmed by the technical discussions replete with professional jargon and attention to methodological detail. This poses a barrier to their becoming meaningfully involved. In one meeting, because of the evident drift of such people toward the exit, an intermission was called, a rump session held, and a decision made to redirect the discussion after reconvening to the interests of the "users." This can be done--and can be helpful--but it runs counter to the normal way that the EMF working group tends to operate.

Analysts, and modelers especially, are solution-directed. They devote themselves primarily to finding ways to deal with problems, rather than in exploring the nature of the problem and making sure that the right questions are being asked and the correct assumptions being made. They will understandably (and usually unconsciously) favor formulation and assumptions for which their methods work best. Their tools can be limiting and their technical skills can become blinders. A modeler, viewing a problem as a control system, for example, may think in terms of "optimal paths," even though this notion may not be especially appropriate in the context being considered.[28]

Policymakers are somewhat similar in this regard. Although they are not typically beholden to any one particular methodology or means of seeking answers, they are (like everyone else) inclined to see things and do things in the manner that feels most comfortable. Yet policymakers have a clear responsibility to identify issues as well as they can. They are readier and better equipped than analysts to handle political and qualitative aspects of a problem (which the analysts may be tempted to assume away), and they have more experience handling people, finding compromises, and balancing the conflicting interests of a highly diversified society. Attention to the question-asking and assumption-setting phase of policy analysis falls principally to the policymaker.

For these reasons, the following suggested steps for making policy models more intelligible and useful to policymakers start not with the policy modeler, but with the policymaker. The goal is first and foremost to involve decisionmakers in policy analysis in a meaningful way and to make policy models easier for those without a technical orientation to use, understand, and accept.

HUMANIZING POLICY ANALYSIS

The initial step is informal and indirect. It seeks to engage policymakers in helping to frame an analysis without their necessarily knowing they are playing that role. This can be accomplished by one-to-one visits, small seminars, and short, simply-worded questionnaires. The focus is on identifying central issues, important points of view, and major points of contention.[29]

The second step is to locate existing or ongoing studies applicable to the issues identified. In evaluating and comparing alternative approaches to a problem and the assumptions made, the work of the Energy Modeling Forum and the MIT Model Assessment Group can be very helpful.

The third step is to rephrase or interpret the reasoning and results of selected studies so as to make the essence of their arguments transparent to the decisionmaker. Policymakers will not be ready to believe answers that run counter to their intuition or policy position if they have no way of checking these answers or of understanding how they were derived.

The typical policy model is not designed for ease of communication with the decisionmaker. Even the model builder may have difficulty comprehending fully the essential workings of the model. Ideally, the model should be presented to the policymaker, not as a "black box" with assumptions and data feeding into the left and results coming out from the right, but as an "open box" whose inner workings are sufficiently simplified, exposed, and elucidated to enable the policymaker to trace the chain of causality from input to output on at least an elementary and fundamental level.

Models are rarely presented as open boxes. It would be a research project of considerable intellectual content and practical significance to develop open-box versions of selected models of greatest potential interest to policymakers. These would be both models that address questions that policymakers are asking and models that would help policymakers determine which questions they should be asking.

An open-box model should be as simple and transparent as possible. For a large regionally disaggregated model, the open-box version might consist of little more than a description of the accounting conventions, component relationships, and linkage mechanism. For a more intricately structured or nested model, the open-box version might be the logical inner kernel of the model--the part that generates the dominant modes of behavior. Once these basic properties are understood, additional layers of complication can be added in a deliberate manner so that the changing character of the model and its implications can be followed.[30]

In presentation before a group of decisionmakers, the open-box strategy would be to start out with the simplest version of an analysis that captures its basics, and use this partly as a straw man to stimulate interest, disagreement, and questioning; then add further elements as the discussion develops and circumstances dictate, always being careful not to go beyond the point of feasible comprehension and effective communication.

This strategy sounds utopian, but a version of it was in fact used very successfully (albeit unintentionally) in the meeting of the EMF senior advisory panel that drew attention to the importance of capital investment in tracing the effects of energy on the economy.[31] Formulating the right questions to ask--rather than finding answers--should be the central objective.

GETTING ORGANIZED

Experience to date is sufficiently encouraging to recommend formation of a group (or groups) of interested executives and policymakers set up specifically to explore and implement the proposed initiatives. Such a group would meet regularly for policy discussion and analysis of timely issues. Call it an Energy Policy Analysis Review (EPAR) group. There are several endeavors with some similarities to EPAR, including the international WAES and WOCOL studies organized by Carroll Wilson,[32,33] activities of the Committee for Economic Development,[34] and the Energy Research Group of utility executives (ERG), which comes together to discuss problems facing the electric utility industry.[35] Critical to the success of such efforts is the skill of the supporting staffs. In the case of ERG, support is provided by a private consulting firm (NERA) working with the utility companies. For the international efforts, staff was made available largely by the members themselves. For EPAR, one would want a staff of creative third-party model analysts with perceptive understanding of the policymaker, as well as the modeler and the methods of analysis.

EPAR would bring selected policy analysts into its discussions as guests and resource lecturers. The staff members of EPAR would work closely with these analysts in advance of a meeting to help them phrase their studies in "open-box" form to facilitate discussion and communication. EPAR could provide an important new communication medium for policy analysts and would give them the kind of valuable exposure and feedback many of them now lack.

Forum analysis and model assessment have so far been paying the largest direct dividends to the technically minded community of modelers and sophisticated model users, who have been learning a great deal about models and the uses of models through this process. One of the lessons learned is that posing policy questions to a model, studying its answers, comparing these answers with those supplied by other models (including simplified and complicated versions of the given model), and understanding the reasons for the differences is a powerful means for learning about the model. It is also a very good way to deepen insights about the policy questions to which the model is being applied.

These insights would be as useful and important to the policymaker as they are to the modelers and analysts, and at least as germane to their work. A significant advance in the usefulness and application of policy analysis will come with the development of means for including the policymaker more naturally in the analytic process. Bringing policy analysis firmly into the mainstream of human affairs should be a first order of business for those concerned about the health of the policy establishment and its ability to cope with world problems. It is a good way to deal with the skepticism about policy analysis--not by public relations but in the only truly effective manner: by constructive action and new initiatives--with policymakers as the initiators.

ACKNOWLEDGEMENTS

This paper has benefited from the suggestions of several friends and colleagues, William Hogan, Alan Manne, and Philip Sisson among them. The author takes full responsibility for the views expressed.

REFERENCES

1. Cameron, J., "The Economic Modelers Vie for Washington's Ear," Fortune, November 20, 1978, pp. 102-104.

2. Greenberger, M., Crenson, M. A., and Crissey, B. L., Models in the Policy Process: Public Decision Making in the Computer Era, New York: Russell Sage Foundation, 1976.

3. Greenberger, M. (ed.), Computers, Communications, and the Public Interest Baltimore: The Johns Hopkins University Press, 1972.

4. Males, R. and Richesl, R., "Economic Value of the Breeder Technology: A Comment on the Ford-MITRE Study," Palo Alto: Electric Power Research Institute, April 12, 1977.

5. "Energy: An Uncertain Future, An Analysis of U. S. and World Energy Projections through 1990," prepared at the request of the Committee on Energy and Natural Resources, U. S. Senate, Washington, D. C.: U. S. Government Printing Office, Publication No. 95-157, December 1978.

6. Cazalet, E. G., Clark, C. E., and Keelin, T. W., "Costs and Benefits of Over/Under Capacity in Electric Power System Planning," Palo Alto: Decision Focus, Incorporated, EPRI EA-927, October 1978.

7. U. S. Nuclear Regulatory Commission, "Nuclear Regulatory Commission Issues Policy Statement on Reactor Safety Study and Review by Lewis Panel," Office of Public Affairs Press Release No. 79-19, Washington, D. C.: January 19, 1979. Also, H. W. Lewis, et al, Risk Assessment Review Report to the U. S. Nuclear Regulatory Commission, NUREG/CR-0400, available from National Technical Information Servbice, Springfield, Virginia, 22161.

8. Testimony of H. W. Lewis before the Subcommittee on Energy and the Environment of the House Committee on Interior and Insular Affairs, U. S. Congress, Washington, D. C., February 26, 1979.

9. Kemeny, J. G., et al., The Need for Change: The Legacy of TMI, Report of the President's Commission on the Accident at Three Mile Island, Washington, D. C.: October 1979.

10. U. S. Energy Research and Development Administration, "A National Plan for Energy Research, Development, and Demonstration: Creating Energy Choices for the Future," ERDA 48, Washington, D. C.: U. S. Government Printing Office, 1975.

11. Personal correspondence with author, 1976 and 1980.

12. Khazzoom, J. D. (ed.), "Workshop on Modeling Interrelationships between the Energy Sector and the General Economy," Palo Alto: Electric Power Research Institute, No. 75-51, 1976.

12a. Hudson, E. A. and Jorgenson, D. W., "U. S. Energy Policy and Economic Growth, 1975-2000," Bell Journal of Economics and Management Science, Vol. 5, No. 2, Autumn 1974, pp. 461-514.

13. A Time to Choose: America's Energy Future, A Report to the Energy Policy Project of the Ford Foundation, Cambridge: Ballinger Press, 1974, pp. 493-511.

14. Holloway, M. L., "The Importance of Transfer in Understanding Large Scale Models," paper presented at the annual meeting of the American Association for the Advancement of Science, San Francisco, January 1980.

14a. The book on the PIES evaluation is being published in two parts. Part I is scheduled for release in the Spring of 1980 by Academic Press. Part II containing 11 detailed studies will be made available as a companion book at the same time by the Texas Energy and Natural Resources Advisory Council.

15. National Advisory Board, Texas National Energy Modeling Program, minutes of meeting, Washington, D. C. June 11-12, 1979.

16. Thrall, R., et al., Transfer and Alternatives to Transfer for Large Scale Models, session at the annual meeting of the American Association for the Advancement of Science, San Francisco, January 5, 1980.

17. Greenberger, M., Richels, R., "Assessing Energy Policy Models; Current State and Future Directions," Annual Review of Energy 4, 1979. Also, Greenberger, M., "A Way of Thinking About Model Analysis," Proc. Workshop on Validation and Assessment Issues of Energy Models, Gaithersburg, Maryland: National Bureau of Standards, 1979.

18. MIT Model Assessment Group, "Independent Assessment of Energy Policy Models: Two Case Studies," Report No. 78-011, Cambridge, Massachusetts: MIT Energy Laboratory and Center for Computational Research in Economics and Management Science, 1978. Also, Model Assessment Laboratory, First Year Report, 1978.

19. Sweeney, J. W. and Weyant, J. P., "The Energy Modeling Forum: Past, Present, and Future," Energy Policy, 1979. Also, Hogan, W. W., "The Energy Modeling Forum: A Communication Bridge," paper presented at the 8th Trienniel IFORS Conference, Toronto, June 20, 1978. Also, Hogan W. W., "Energy Models: Building Understanding for Better Use," paper presented at the Second Lawrence Symposium on Systems and Decision Sciences, Berkeley, October 3-4, 1978. Also, Koopmans, T. C., et al., "Energy Modeling for an Uncertain Future," Report of the Modeling Resources Group, Synthesis Panel of the Committee on Nuclear and Alternative Energy Systems, National Research Council, Washington, D. C.: 1978.

20. Energy Modeling Forum, Energy and the Economy, EMF 1, Stanford University, Stanford: September 1977.

21. Energy Modeling Forum, Coal in Transition: 1980-2000, EMF 2, Stanford University, Stanford: July 1978.

22. Energy Modeling Forum, Electric Load Forecasting: Probing the Issues with Models, EMF 3, Stanford University, Stanford: April 1979.

23. Energy Modeling Forum, "Aggregate Elasticity of Energy Demand," Working Paper, EMF 4.4, Stanford University, Stanford: January 1980.

24. Energy Modeling Forum, "U. S. Oil and Gas Supply Summary Report," Working Paper, EMF 5.4, Stanford University, Stanford: November 1979.

25. Energy Modeling Forum, World Oil: Supply, Demand, and Price, Agenda for Study, Stanford University, Stanford: January 1980.

26. Greenberger, M., "Closing the Circuit Between Modelers and Decision Makers," EPRI Journal, Palo Alto: Electric Power Research Institute, October 6-13, 1977. Also, NATO Advanced Research Institute, Proceedings on the Application of Systems Science to National Energy Policy Planning, Report of Working Group 4, "The Communication Problem in Energy Policy Analysis," Islip, New York: Brookhaven National Laboratory, November 1979.

27. Peck, S., "Six Bridges Between the Developers and Users of Energy Models," Proceedings, Second IGT Symposium, 1979. Also, Greenberger, M., "Closing the Circuit Between Modelers and Decision Makers," EPRI Journal, Palo Alto: Electric Power Research Institute, October 3-6, 1977.

28. The concept of an "optimal path" is adopted in much of the economic modeling work on the world price of oil. See, for example, Hnyilicza, E., and Pindyck, R. S., "Pricing Policies for a Two-Part Exhaustible Resource Cartel: The Case of OPEC," European Economic Review, 8, pp. 139-154, Cambridge, Massachusetts: M.I.T., 1976. Critics of such work say that petroleum exporting countries are not making their decisions so as to optimize their future stream of revenues. Political factors and short-term domestic requirements are seen by these critics as dominant in determining pricing and production behavior.

29. Points of contention are discussed in Crissey, B. L., "A Rational Framework for the Use of Computer Simulation Models in a Policy Context," Ph.D. Dissertation, Baltimore: The Johns Hopkins University, November 1975. The problem of identifying important issues for consideration by decision makers is treated in Greenberger, M., "Improving the Analysis of International Energy Issues," Report to the Rockefeller Foundation, New York: 1979.

30. An analogy to the strategy of gradually increasing complexity is found in applied mathematics in the development of continuation methods. See Kolata, G. B., "Continuation Methods: New Ways to Solve Equations," Science 204, pp. 488-489, May 4, 1979.

31. The "straw man" that stimulated discussion in the EMF senior advisory panel meeting was from Hogan, W. W., and Manne, A. S., "Energy Economy Interactions: The Fable of the Elephant and the Rabbit?," in Hitch, C. J. (ed.), Modeling Energy-Economy Interactions: Five Approaches, Resources for the Future, Washington, D. C.: 1977.

32. Wilson, C. L., et al., Workshop on Alternative Energy Strategies (WAES), Energy: Global Prospects 1985-2000, New York: McGraw Hill, 1977.

33. Wilson, C. L., et al., World Coal Study (WOCOL): Global Perspectives to 2000, Cambridge, Massachusetts: MIT, November 1979.

34. The Committee for Economic Development convenes world leaders to consider international problems of timely interest. For an example of a CED report, see "Redefining Government's Role in the Market System," New York: Committee for Economic Development, July 1979.

35. The Energy Research Group was organized by Irwin Stelzer of National Economic Research Associates, Inc. (NERA). NERA performs studies and provides discussion papers for consideration by the ERG group. See, for example, the Executive Summary to the Report on "The Impact of Alternative Policy Reactions to Three Mile Island," National Impacts, Volume 1, New York: June 12, 1979.

REMARKS ON WOOD AND GREENBERGER PAPERS

KEN HOFFMAN, SENIOR VICE PRESIDENT

MATHTECH, Inc.

There are many useful insights in both of these papers, and I could spend a couple of hours agreeing with most of the material covered. I think I would rather address some of the more controversial statements and the areas of disagreement that may exist. There were several statements that were designed to be provocative, to get us to think about and discuss the issues in greater depth.

Starting with Dave Wood's paper, I think his discussion of measures of model confidence gave us a good state-of-the-art review of where things stand on rating documentation, verification, validation, and the usability of models. He mentioned the difficulty in applying these criteria to the steps of verification and validation. I feel strongly that one of the contributors to that difficulty is our failure to make a sharp distinction between the purposes that different models serve.

Some applications require normative or prescriptive models that do not claim to simulate all of the problems or market imperfections that can arise, but look at an idealized solution to a policy or an idealized implementation of some technology. This kind of modeling is something like the Carnot cycle analysis in thermodynamics, where one is not concerned about the practicalities of friction and irreversabilities that arise in the process, but is concerned about what is the best one could do in an idealized system. I think that a lot of the normative and prescriptive modeling is of that classification and gives much useful insight. It may not be a good simulation, but it does give a lot of useful information concerning which direction one might move in.

Going down through the purposes of modeling, there are some models that are purely descriptive in nature. They are meant simply to make sure that known relationships are captured and are reflected in the analysis. This is an important feature, but we should not expect such models to provide forecasts or to give good normative perspectives on the problem; they are more like the design methods used in engineering.

There is also the direct simulation model that attempts to deal with a very complex set of interrelationships for forecasting purposes, or perhaps for more narrow purposes. This type of model requires a very different set of validation and verification criteria.

I think we have to keep these different purposes of modeling in mind when we ask the validation and verification questions. Such discipline would go a long way toward organizing classes of questions that one might ask and perhaps make it a bit easier to deal with these measures of model confidence.

I believe that all of these classes of modeling are very important, but often there is competition and the viewpoint that only one of these classes ought to be done and not the other. In some of the models we often mix the approaches used--for example, taking a normative approach and mixing it with a behavioral simulation approach--and we end up with something that is trying to cover both of these aspects and really does not do any one very well.

I think the sociology of modeling, as described by Martin Greenberger, is very interesting. A historical record should be maintained and there should be some place to which everyone can send their latest experiences and stories to be kept on file so that we all can learn from them. We can learn a lot from the problems of misapplication and misinterpretation of models, and retrospective analysis of their usefulness.

I do want to take strong disagreement with one of Martin Greenberger's points dealing with the differences between policy modeling and science. I think such differences may have existed in the past, but if we allow these differences to persist we are losing a lot of the strengths and experience that we have learned over a long period of time from applications of the scientific method. I do not think we should let go of that connection and lose all of the difficulties that science has suffered in poor documentation and poor verification of scientific results. We must look into the things we have learned from applications of the scientific method and even bend our own perspectives and viewpoints on modeling and policy analyses to make them fit the scientific paradigm. If we let go of that connection and say that policy analysis is very different, I think we loose too much and are destined to suffer the same mistakes that have been made throughout the years in the field of science.

To clarify my viewpoint, what we are dealing with in both policy analysis and in the more traditional technical fields is the application of knowledge and basic information by users or decision makers. This is not a new problem. In science we are looking to learn basic information--the fundamentals that apply to a technical problem. There is a whole group of people between science and its application who traditionally have applied science to real problems for real users; that is the engineering profession. We can learn a lot from the way the engineering profession fulfills its role in the application of science to real world

problems, to provide that same bridge between the basic data and theory of models and their application to real policy issues. In policy analysis we too often blur the distinction between those people who seek basic information and theory, and those who apply that information to problems for users.

Dave Wood paraphrased a debate that went on at the last meeting of this group. There was apparently one claim that we should not go beyond the limits of science in answering questions that policy makers ask, while the other position claimed the questions always are more complicated than the information we have to deal with, and that we must invariably stretch beyond the available information. In fact, in the real applications of science, we have always stretched beyond the available information. Bridges were built and structures were built before the nature of their materials was really understood. Things like "fudge" factors, ignorance factors, safety factors, and fail-safe analyses all came into the terminology and into practical use. I think these same mechanisms and procedures are very appropriate in policy analysis. Decisions are going to have to be made that stretch well beyond the information that is available to us. The practical problem, then, is how to provide analysis relevant to such decisions and arrive at policies that are fail safe and have proper ignorance factors built into them. That is the challenge, and it is a very different function from the science of modeling.

Talking about distinctions being blurred, I am reminded that most of those here are, or have been, model builders. I used to be a model builder; I now consider mysef to be a model user. Considering the research emphasis that is being placed on evaluation, I guess I ought to become a model evaluator. We must remember, however, that these are three very different functions. We have examples from other areas. In the arts, for example, we have performers, we have the audience, and we have the critics. It is very rare that you find someone who is all three, or even two out of the three, because of obvious conflicts of interest and the loss of credibility.

It seems that in modeling we very frequently find people who are all three, or who claim to be all three. Can we let people get away with that, or should there be a more definitive statement about what that individual is really trying to accomplish in his or her research, and about the role that he or she is really playing?

I have two more generic issues that I would like to raise. In these discussions we have heard some good examples of what has gone on in the past, and we can learn much from those examples. I think it would also be interesting to look at some obvious needs and gaps in what has gone on over the last few years and see if we can learn from these situations as well. Everyone will

have his favorite idea of a gap, or of something that has not been dealt with particularly well in modeling, analysis, or in the gathering of basic information.

One of my high priority problems is the important need for a technology data base. Many of the mid-term and long-term models rely on some characterization or description of technologies. It turns out that when you evaluate each model, each contains its own technology data base, and I think it would be a great step forward to at least come up with some uniform technological descriptions that are related to current and planned research and development programs. The characterization of emerging technologies, I believe, is an important piece of information and ought to be dealt with.

Another critically important information need deals with end use stock and turnover and its relation to energy demand analysis. We have read that energy demand turned down in 1979. Is that attributed to changes in economic structure, or can it be attributed to a response of consumers to install more efficient devices and to use the devices that are currently installed more efficiently? I think the whole area of end use data and information needs further development, and this topic would be worth some discussion.

The last problem area that I want to mention deals with regional analysis and forecasting. We have seen some progress and interaction between Federal level analysis and regional analysis. I think that this workshop should explore how that could be strengthened, and how that work should proceed in a way that meets some of the measures of confidence that have been outlined.

The last formal comment I have on the two papers presented in this session deals with the proper utilization of the creative talent that is assembled here to discuss validation and verification questions. We see a situation where the large institutionalized models and the essentially defensive questions of validation and verification seem to be soaking up a substantial portion of the funds available in the energy modeling and analysis field. Any time that I want to feel depressed, I look at the amount of basic and applied research money devoted to new modeling approaches. We <u>know</u> that the approaches we have now overemphazize perfect markets and perfect decisions. Others take different approaches, but we <u>know</u> none of them are really entirely valid for good forecasting or decision making. What are we now doing about the next generation of models to improve the situation? I think this is a challenge to a workshop like this: out of this workshop and the overall validation process should come some research directions in energy analysis. I hope that this need is not lost in the discussion and in the results of the proceedings.

I have seen some interesting new things lately, like applications of game theory and some further analysis with nonequilibrium economics. We held a workshop at Brookhaven on world oil pricing about three years ago and had to look awfully hard to find someone who was looking at nonequilibrium behavior to oil pricing, even though we had excellent examples that indicated this mode of decision making. Someone called the work of Dermot Gately at New York University to our attention. He is an economist who produced curves and future projections of prices that have discontinuities and jumps, and all of those characteristics of nonequilibrium behavior. That field has grown since, but I think there is even more to be done there.

I should finish up with the story behind that ERDA analysis that Martin referred to and asked me to address. It is described very well in Martin's paper, but since it takes nine to twelve months to get these papers out, I think I should discuss it now and not leave the record with a void on this issue.

The claim in the mild controversy following release of the ERDA plan was that conservation was overlooked in the supporting analysis and that it was the fault of the model used. Further, it was said that this led to an underemphasis on conservation in the R&D plan.

In fact, the model used was really the first one to include end use devices and the end use part of the system at the same level of detail as the supply technologies. Further, the conservation scenario described in an Appendix to the ERDA R&D plan had the largest beneficial impact in terms of environmental emissions, cost, capital cost, and resource utilization. It looked like a real winner.

The decision on whether more R&D would be sponsored in those areas, and to what extent, was the next question, but a somewhat different one. As I understand that policy decision, given all the information that came out of the analysis on the value of conservation, it was felt that the actual R&D was the responsibility of the private sector. The private sector built the furnaces, automobiles, heat exchangers, and other end use devices, and it was probably felt that this area was much too consumer oriented and not the responsibility of the government. The answer then, or the decision, was that conservation R&D was not the highest priority area for government funds, but was a matter to be dealt with through standards. In a subsequent plan, conservation R&D was raised to the highest priority; however, the debate over what is an appropriate government role in conservation continues today. Even now, pricing mechanisms and standards are given the primary role in encouraging conservation.

To return to the issue of possible misapplication of a model, I would recommend you read the section dealing with that analysis and the subsequent debate. There is a lot that may be learned from that case, as well as the others Martin has described.

BUILDING GOOD MODELS IS NOT ENOUGH

Richard Richels
Electric Power Research Institute

The potential advantages of policy modeling are many. The very discipline of constructing a model can help us get our thinking straight. Models can provide a bookkeeping mechanism, tell us what kinds of information are desirable and when we need it, allow for the possibility of experimenting with the model rather than the system itself, and facilitate communication among those concerned with a policy issue. Perhaps, most importantly, models can have an enormous educational value, helping us develop general insights into a problem that can serve as a basis for the intelligent and informed making of decisions (1). Yet, in the case of energy policy modeling, many of these advantages have not been realized. As a group of researchers on policy modeling observed, "...there is one thing that policy models have in common. Most fall short of their potential as instruments for the clarification of policy issues and the enlightenment of policymakers" (2). The failure of models to live up to the expectations and demands of the user community has led to increasing skepticism from policymakers about the usefulness of policy modeling. This paper examines two reasons for this growing dissatisfaction and considers ways to deal with it.

Sources of the Dissatisfaction

General Glen Kent, formerly head of studies and analysis for the U.S. Air Force, noted several years ago that "decisionmakers are becoming increasingly annoyed that different analysts get quite different answers to the same problem" (3). When this happens, it is natural to want to take a closer look at the models and find out more about how they come to produce such puzzling results. Until recently, however, there was very little evaluative information available concerning the quality and appropriate uses of individual models. Policymakers were forced to rely almost entirely on the experience and judgment of the model builder for an evaluation of the strengths and weaknesses of a particular model. As model users have become more sophisticated they have begun to question the ability of the model builder to be an objective critic of the fruits of his own labor. The lack of independent evaluative information has led some to worry that modelers may be obscuring deficiencies of major concern in model application. The problem has become compounded as the number of models has increased and policymakers have become deluged with conflicting claims from model developers anxious to convince users and would-be users of the value of their creations.

Users have also found themselves thrust into the uncomfortable position of model defender. Once a model begins to be used as an

instrument of debate, it will be seen by some as a threat or challenge and become the subject of investigation (4). Typical of such challenges is the American Public Gas Association's attempt to discredit the Federal Power Commission's use of a model in establishing natural gas rates. APGA argued that "the 'economic models' dreamed up by producer-sponsored consultants and untested by cross-examination, do not begin to rise to the status of 'substantial evidence'" (5). The FPC found itself in the position of having to defend its choice of models in court.

Such situations have led to calls from the user community for independent evaluative information regarding the overall quality and applicability of energy policy models. As a result, organizations that have been sponsoring model development, such as the Department of Energy and the Electric Power Research Institute, have begun shifting substantial resources to assessment activities. There has been a growing realization that independent or third-party assessment is critical if policymakers are to have confidence in model results. The nature of these activities will be discussed in the next section.

But quality control is only part of the problem. Stokey and Zeckhauser, in their primer on policy modeling, caution the aspiring modeler that "the world will never beat a pathway to your door just because you build a better model; analysis is worthless if it can't be communicated to others" (1). Even with excellent communication it is hard to imagine anyone beating a path to a modeler's door; nevertheless, the above quotation identifies a basic weakness with policy analyses. Each year, countless studies, many containing useful and timely information, gather dust (6). Sometimes the fault is with the policymaker for failing to take advantage of readily accessible information. But more often the blame is with the modeler. Analyses are presented in such a way that they are more likely to confuse and overwhelm than inform. This failure to communicate seriously impedes the process of providing policymakers with important tools for policy making. The result is a situation where analyses are used more to <u>rationalize</u> policy decisions than to serve as a basis for <u>rational</u> decisionmaking.

Policy modeling can be enormously valuable in providing general insights into a problem. Unfortunately, the educational benefits up to now have accrued almost entirely to the modelers and have failed to reach the policy arena where they are most needed. This is a major weakness with policy modeling. Hogan hit the mark when he wrote that "the purpose of energy modeling is insight, not numbers" (7). <u>The problem is that policymakers are being presented with numbers, not insights</u>. Without the rationale to support these numbers, it is small wonder that they have little use for results that do not conform with their own intuition. If models are to reach their full potential, more attention must be placed on communicating the rationale of policy models and not just the results.

Independent Model Assessment

The first step to making policy models more useful in policy has already been taken. The last few years have seen tremendous growth in the area of independent model assessment. A number of assessment projects have been established and are providing valuable information essential for the intelligent use of models (4). The work done by the Energy Model Analysis Program at the Massachusetts Institute of Technology is an example of this model oriented type of analysis (8).

With third-party assessment, the work is done by individuals independent, organizationally, of both model developers and users. The evaluation responsibilities do not fall on the modeler whose incentives are chiefly to get the model working and used. Nor do they fall on the model sponsors, who may have a sense of commitment to the model through their association with the developers (4). Greenberger *et al.* describes the model analyzer as a new breed of researcher/pragmatist--"a highly skilled professional and astute practitioner of third-party model analysis" (2).

In a typical evaluation, the model is examined over the full range of its capability. The assessors begin by going through the computer code line by line to verify that the model is as advertized. The validity of the model is then examined in the context of the purposes for which the model was constructed or is to be used. The assessors attempt through a series of validity checks to assess the agreement between the behavior of the model and the real world system being modeled (9).

Verification and validation are intended to give the user some indication of the "usefulness" of a model. The assessors are also concerned with the "usability" of the model. "A model is usable if it is understandable and plausible to others than its developers, economic to run on a computer, and accessible to those who wish to use it (10)." The usability of a model depends on such factors as the quality of its documentation, operating characteristics and overall portability--all of which are assessed by the model analyzers.

The best time for third-party evaluation is an open question. Independent assessment can and does take place at each stage in the modeling process--at the creation stage, at the application stage, and even as the model is being used as an instrument of debate. What is clear, though, is that if careful examination is put off too long, the results are apt to be disappointing. Model users need evaluative information before selecting a model for a policy debate. Without the careful scrutiny of highly trained analysts, users can only speculate about a model's capabilities, limitations, adequacy and realism. The likelihood of a poor choice is high. As Figure 1 suggests, independent assessment should first occur early on in the modeling process, before the model is brought to the policy arena.

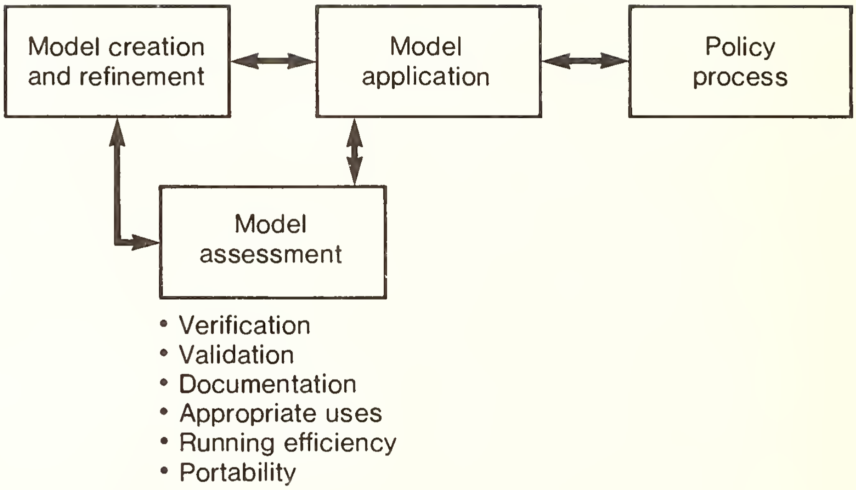


Figure 1. Model Assessment and the Policy Process

Independent assessment should not be considered a "one-shot" activity. An active model is in a constant state of evolution. As it is brought to new applications it is modified and refined. Careful examination by third-parties also leads to modifications as the model builder is provided with valuable feed-back signals helpful in correcting and improving the model. If an assessment is to remain timely it must not ignore the changing nature of the model. Frequent assessment "updates" will be necessary as long as the model remains active. For this reason model assessment is depicted as an iterative activity in Figure 1.

A great deal has been learned about the nature, problems, and possibilities of independent model assessment over the last few years. Activities such as those underway at MIT have reinforced the conviction that third-party assessment is essential if policymakers are to begin to have confidence in policy models (8). But model assessment is only part of the answer. Experience has shown that even when a model is perceived as "useful" and "usable," it still may not be "used." This brings us to the second reason why models have failed to live up to the hopes and expectations of the user community; the model builder is not communicating effectively to decisionmakers the insights, structure, and understanding available from the model. In the next section, we look at ways to bridge the "communication gap" between the model builder and user.

User Analysis

The purpose of policy analysis is to help policymakers make better decisions. Decisionmaking can be aided by using the model as a tool for education and exploration. There is considerable evidence that models can be effectively used to improve understanding. Examples abound as to how a model result which initially appeared anomalous and implausible led to reassessment of ideas and a deepening of insights (11). Modelers encounter such instances daily. The problem is in making the insights available to those responsible for policymaking. It is not sufficient to produce a technical report only understandable to the trained analyst. A primary reason why policy analyses go unread is that they are too long or too hard to understand by the lay reader (1). Converting technical results into a language comprehensible to policymakers is essential if model results are to be used.

The difficulty with bringing together people with such diversity in backgrounds as modelers and policymakers is illustrated by the following amusing anecdote:

> A congressman who had been persuaded to favor a "soft-technology" approach to the energy needs of the United States was discussing projections of demand for electricity with a modeler. He pointed out that with adequate conservation measures, a little modesty in our life-style, and other steps

> emphasized by the soft-technologists, the rate of growth of electrical demand could easily be held down to 2 percent per year. "But Mr. Congressman, even at the low rate of 2 percent per year, electrical demand will double in thirty-five years," said the modeler. "That's your opinion!" replied the Congressman (12).

Although the story may seem somewhat extreme, it is indicative of a very serious communication gulf. To bridge this gap, it may be necessary to add a user oriented activity to our earlier diagram. To connote the intended beneficiary, we will refer to it as "user analysis." Whereas model assessment focuses primarily on quality-related issues, the emphasis here is on transferring the insights of a policy analysis to the policymaker. The intention is to insure that the communication of a policy analysis receives as much attention as its design and implementation.

User analysis, as portrayed on the right side of Figure 2, has several components. The first, the presentation of results in the context of the policy problem under study, may seem obvious, but it is all too often overlooked. The modeler should highlight those aspects of his model or analysis that are relevant to the decision problem of interest to the policymaker. To do this he may wish to employ the techniques of decision analysis, a methodology designed specifically for aiding decisionmaking (13). Decision analysis forces the analyst to focus on the decision. The heart of the methodology is the decision tree which lays out the relevant sequences of decisions and outcomes. For example, the decision tree of Figure 3 was used by the Synfuels Interagency Task Force in 1975 to examine alternatives for implementing a synthetic fuel program in the U.S. (14). The squares denote decision nodes and the circles denote chance nodes. The initial decision was a choice among four alternative government financed synthetic fuel programs.

The advantage of a decision tree is that it focuses on the sequential nature of the decision problem. In structuring alternative strategies in a decision tree framework, the analyst works with the decisionmaker to identify those decisions that must be made today and separates them from those which can wait and consequently benefit from the inflow of additional information. In the synfuel analysis, the decision was not whether the U.S. should commit itself to full scale synfuel production, but one of the size of the initial program. Prior to the work of the Synfuel Task Force, the debate had focused on the desirability of a full scale synfuel industry. The work of the Synfuel Task Force is credited with focusing the Government debate on the real decision at hand. The analysis was regarded as a key factor in persuading the Ford Administration to cut back from the President's original goal of a million barrels down to 350,000 barrels a day in 1985 (15).

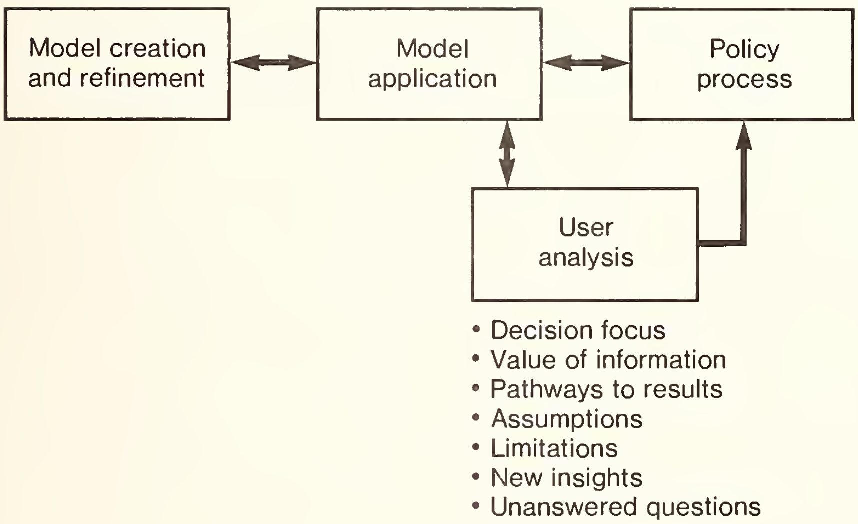


Figure 2. User Analysis and the Policy Process

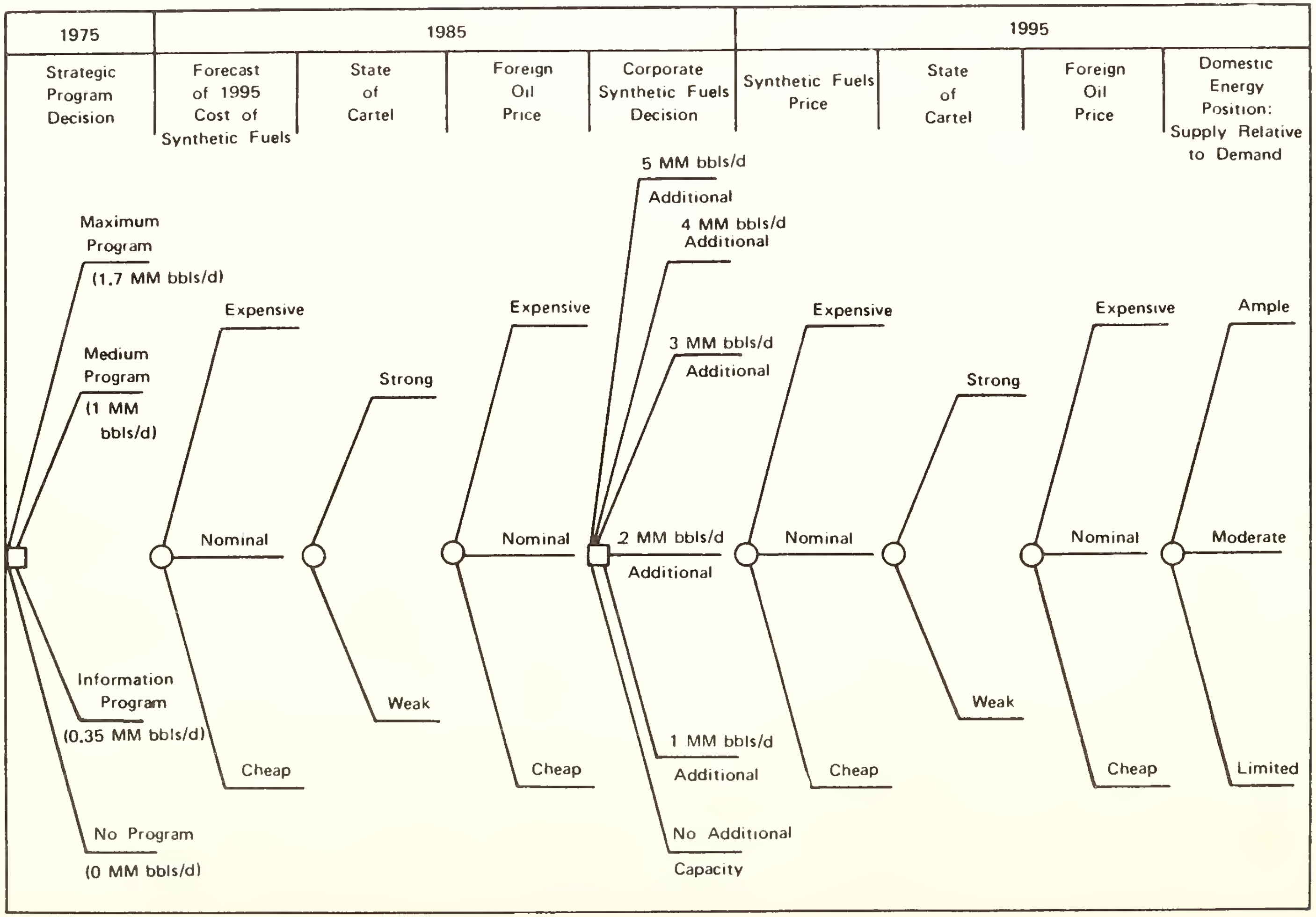


Figure 3. Synthetic Fuels Decision Tree

In presenting the results of a policy analysis, the modeler should also indicate the sensitivity of the decision being studied to the basic assumptions underlying the analysis. In the synfuel decision analysis, the Task Force assumed that the probability of a strong oil producers' cartel in 1985 is 50 percent. Figure 4 shows the sensitivity of the optimal program to this assumption (15). If the decisionmaker feels that the probability of a strong cartel in 1985 exceeds 76 percent, the expected net benefits of a synfuel program become positive. Sensitivity analysis provides the policymaker with valuable insight into the relative importance of the uncertainties complicating his decision. If it turns out that a decision is very sensitive to a particular variable, the value of better information regarding that variable is high, and the policymaker may wish to allocate resources to reducing the level of uncertainty.

When discussing the value of information it is also useful to give some indication as to how accurate information must be, to be of value. For example, it is well known that assumptions about uranium resources can affect the desirability of advanced nuclear technologies. For this reason, the government is spending millions of dollars to reduce uncertainty regarding the uranium resource base. Recent analysis has shown, however, that attaining the large potential benefits from better information will be difficult, since the information must be extraordinarily accurate to have a significant impact on the choice of nuclear policy options (16). Such insight can be helpful in determining the potential value of information gathering efforts.

Often model results which appear puzzling at first glance can lead to a reassessment of ideas and a deepening of insights. The communication of interesting results is the modelers biggest challenge. To do this it is not only necessary to present the results, but also the "pathways" to results. At this stage in the presentation, the original model moves to the background and the focus is on telling an intelligible, intuitive story--one that stands apart from the model. Sometimes the modeler may find it useful to rely on a simplified version of the original model to tell the story. The Hogan-Manne fable of the elephant and the rabbit is one such example (17). A simple highly aggregated model was used to illustrate the key concepts that determine the economic impacts of energy policies. The simple model which abstracted from the class of large energy economic models was used quite effectively in communicating the significant energy-economic interactions to policymakers.

In describing the pathways to results, it is essential to translate from specialized modeling jargon into a form that is accessible to a wide audience of potential users. Once the essence of the story is conveyed as clearly and concisely as possible, the modeler can then start adding to its complexity by injecting the necessary caveats about critical assumptions and model limitations and explaining how they may affect the conclusions.

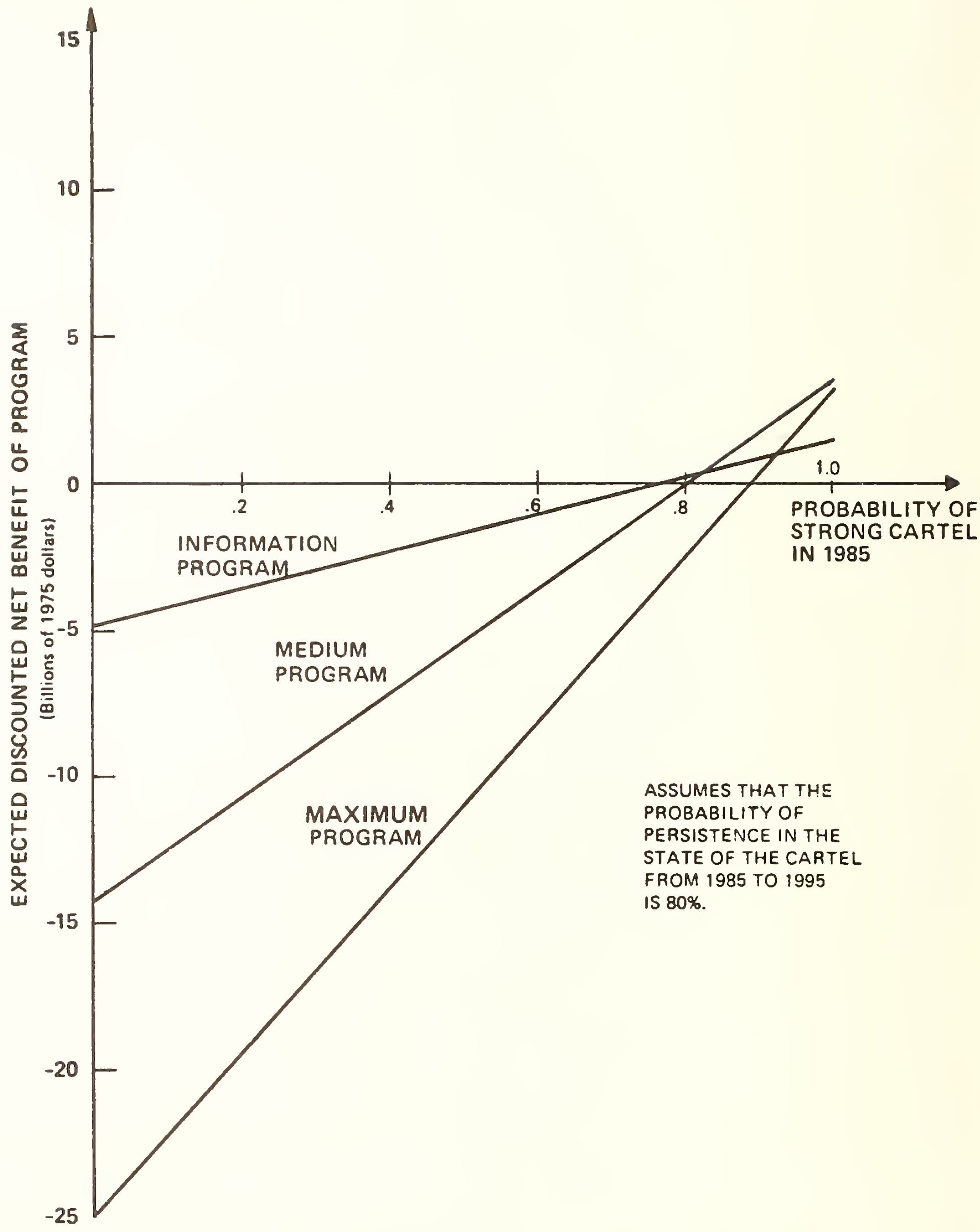


Figure 4. Sensitivity of Expected Net Benefit to the Probability of a Strong Cartel in 1985

This is what Greenberger had in mind when he proposed an "open box" approach to communicating model results to policymakers:

> The typical policy model is not designed for ease of communication with the decisionmaker. Even the model builder may have difficulty comprehending fully the essential workings of the model. Ideally, the model should be presented to the policymaker, not as a "black box" with assumptions and data feeding into the left and results coming out from the right, but as an "open box" whose inner workings are sufficiently simplified, exposed, and elucidated to enable the policymaker to trace the chain of causality from input to output on at least an elementary fundamental level (18).

He goes on to point out that "it would be a research project of considerable intellectual content and practical significance to develop open box versions of selected models of greatest potential interest to policymakers" (18). It would also be a research project of enormous challenge. The true art to model building has always been in keeping the model as simple as possible while still capturing the essence of the system under study. The additional constraint of not exceeding the policymakers comprehension may make the building of the simple model more intellectually taxing than the building of the more intricately structured model from which it is derived. Nonetheless, an open box strategy offers the modeler a vehicle for making the essence of his argument transparent to the policymaker. If such an approach could be successfully implemented, the payoff in terms of more effective communication would more than justify the investment.

The final component to user analysis, identification of important unanswered questions, should define the modeler's agenda for new research. User analysis like model assessment is an iterative activity. It is rare that the resolution of one set of issues does not lead to a whole set of new ones. If the debate is ongoing, the modeler will have a new opportunity to aid decisionmakers. In fact, if he did a good job the first time around, he may even have a mandate for new research!

Final Comments

Figure 5 shows the two kinds of model analysis described above, their relationship to each other, and to the various stages in the modeling process. Model assessment and user analysis are two very different types of activities. Whereas model assessment is best carried out by individuals independent organizationally of both model builders and model users, user analysis should be done as a normal part of the modeler's work assignment. To allocate the activity to third parties is not only inefficient, it breaks the vital link between modeler and policymaker.

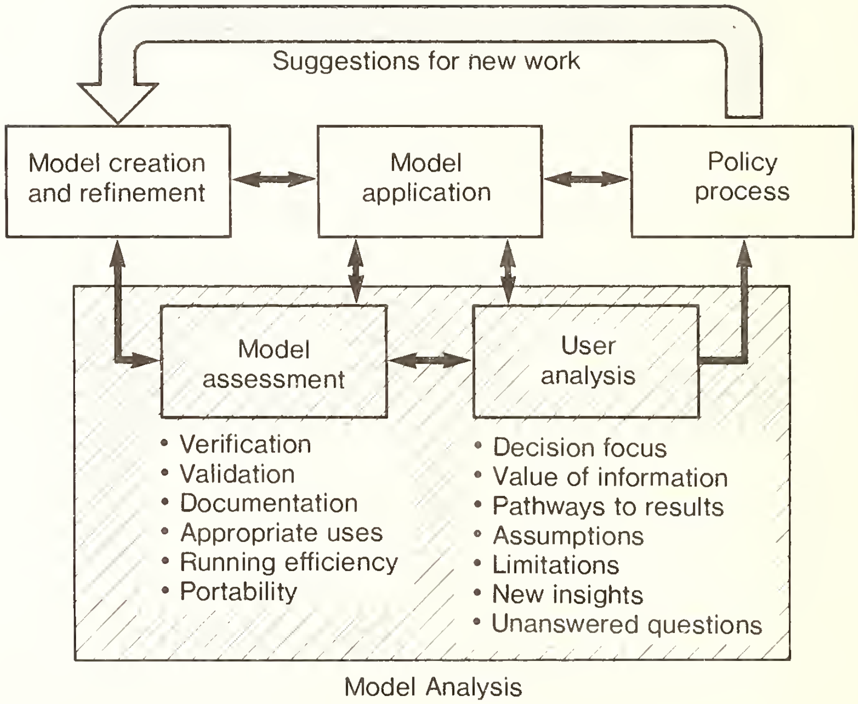


Figure 5. Model Analysis: A Necessaary Bridge Between the Modeler & Policy Maker

With the support and encouragement of model sponsors and users, these activities can become permanent and important parts of the policy process. In the case of model assessment, support is needed for the institutionalization of a new professional discipline. More laboratories and centers are needed where model analysts can come together to practice the art and science of third party model analysis. Ideally, independent model assessment will be set up as recurring activities and the continuing work of permanently funded facilities.

In the case of user analysis, no new institutional framework is called for. Nor do we need to establish a new professional discipline. What is needed is more professional discipline. Modelers must pay more attention to converting technical results into language comprehensible to policymakers. They must put more effort into translating from the specialized jargon of economics and operations research into a form that is meaningful to the nonmodeler. This will not happen, however, until the incentives that the modeler faces reward effective communication. Understandably, modelers do not receive the same feelings of elation and accomplishment in writing a report that come with a model's finally producing plausible results. Sponsors must provide the motivation for more effective communication. From the outset of a study, they must be specific as to what is expected and should frequently review whether standards are being met. "If helpful guidance is what policymakers desire from the modeling project, then helpful guidance is what they must supply to the modelers during the conduct of the effort (2)."

Acknowledgment

I am grateful to Martin Greenberger, Stephen Peck and Stan Sussman for helpful comments and suggestions. The faults and omissions are my responsibility.

REFERENCES

1. Stokey, E., Zeckhauser, R. 1978. A Primer for Policy Analysis. New York: W. W. Norton & Company.

2. Greenberger, M. Crenson, M., Crissey, B. 1976. Models in the Policy Process. New York: Russell Sage Foundation.

3. Brewer, G. 1980. "On Duplicity in the Modeling of Public Problems." Simulation. April 1980.

4. Greenberger, M., Richels, R. 1979. Assessing Energy Policy Models: Current State and Future Directions. Annual Revenue of Energy. 4:467-500.

5. United States Court of Appeals. 1977. The Second National Natural Gas Rate Cases. American Public Gas Association, et al., Petitioners v. Federal Power Commission, Respondent. No. 76-2000, et al.

6. Fromm, G., Hamilton, W., Hamilton, D. 1974. "Federally Sponsored Mathematical Models: Survey and Analysis." Washington, D.C.: Report to the National Science Foundation.

7. Hogan, W. 1978. "Energy Models: Building Understanding for Better Use." Presented at 2nd Lawrence Symposium on Systems and Decision Sciences. Berkeley, Calif., Oct. 1978.

8. Massachusetts Institute of Technology, Energy Laboratory 1978. Independent Assessment of Energy Policy Models: Two Case Studies. Rep. No. 78-011. Cambridge, Mass: MIT Energy Lab.

9. Gass, S. 1977. "Evaluation of Complex Models." Computers and Operations Research 4:27-35.

10. Committee on Merchant Marine and Fisheries 1975. "Computer Simulation Methods to Aid National Growth Policy." 94th Congress, Washington, D.C.: GPO.

11. Hogan, W. 1978. "The Role of Models in Energy Information Activities." Energy Information. Proceedings of a Workshop held at Stanford University, Stanford, CA. Dec. 1977.

12. Sachs, R. 1980. "National Energy Policy - A View from the Underside." National Energy Issues - How Do We Decide. Cambridge, Mass.: Ballinger Publishing Company.

13. Raiffa, H. 1968. Decision Analysis. Reading, Mass: Addison-Wesley.

14. Synfuels Interagency Task Force. 1975. "Recommendations for a Synthetic Fuels Commercialization Program." Washington, D.C.: GPO.

15. Weyant, J. 1978. "Energy Modeling in the Policy Process: Oil Price Decontrol and Synthetic Fuels Commercialization." Presented at 2nd Lawrence Symposium on Systems and Decision Sciences. Berkeley, Calif. Oct. 1978.

16. Gilbert, R., Richels, R. 1980. "Reducing Uranium Resourse Uncertainty: Is It Worth the Cost?" Energy Modeling III: Dealing With Energy Uncertainty, Chicago, Ill. Institute of Gas Technology. Forthcoming.

17. Energy Modeling Forum. 1977. "Energy and the Economy." EMF Report 1. Stanford University, Stanford, CA.

18. Greenberger, M. 1980. "Humanizing Policy Analysis: Confronting the Paradox in Energy Policy Modeling." Presented at the 1980 DOE/NBS Symposium on Validation and Assessment of Energy Models. Gaithersburg, MD. May 1980.

Quality Control for Analysis

George M. Lady[1]

INTRODUCTION

This paper is written as a sequel to that written for the first Department of Energy/National Bureau of Standards symposium on model assessment and validation.[2] Not only is the passage of time and events since that first symposium an advantage, but this paper is being written late in the time allowed for preparing these Proceedings. As a result, I can cite a recent report prepared at the Oak Ridge National Laboratory [2], which provides a sufficient bibliography of papers on model quality control and related topics. I refer the reader to this bibliography rather than reiterate many of those references here.

It should be immediately acknowledged that analysis quality control activities within the Energy Information Administration (EIA) were (and are) also sponsored by the Office of Energy Information Validation, an organization parallel to the Office of Applied Analysis. These activities will not be discussed here. This omission should not be taken in any respect to indicate the relevance or success of the two programs. Instead, this emphasis simply reflects the author's personal experience; further, it represents an account of activities within the modeling enterprise itself.[3] This latter circumstance involves both advantages and disadvantages.

Quality control initiatives proposed (organizationally speaking) within the same budget unit as the associated analysis must survive an essentially continuing evaluation of costs and benefits in relationship to the overall goal of producing analysis results. Accordingly, "more" quality control has an immediate opportunity cost which entails "less" product. It is easy to imagine that the managers who are evaluated in terms of their evident productivity (i.e., quantity and timeliness) do not always accept that measures which reduce this productivity well represent their interests. I consider this circumstance an advantage. The requirement to be specific about the relative value of quality control measures is a useful and proper discipline.

The disadvantages of this arrangement are no more complicated than the problems of myopic planning. A concern over practicality in the heat of decision tends to dramatically discount future benefits.

Resources proposed for expenditure now, but which provide no benefits now, or soon ("soon" sometimes means very soon), might easily be perceived as better spent some other way. The problem of a proper planning horizon pervades resource allocation problems in general and I will not contend that it has been atypically severe for model quality control planning.

The purpose of this paper is to report upon the Department of Energy's program for analysis quality control as developed within EIA's Office of Applied Analysis. This report will not be a detailed accounting of program components, activities, and expenditures, rather, a general statement of issues, the manner of their pursuit and the kinds of problems and outcomes which emerged. These Proceedings and those of the first symposium can leave the impression that the analysis quality control "problem" is to be found in a host of technical issues and associated uncertainties about process... What is to be done? Who is to do it? Who is it to be done for? What form should it take? My intention in writing this paper is to take sharp exception with this view of the "problem." While many issues remain to be resolved, based upon my experience with the EIA quality control program, it is my impression that the quality control problem is in the first order a deficiency in the professional perception of good practice. To be clear, there are many quality control measures which are beyond dispute, but which are not undertaken because they are not professionally satisfying and compete for resources which might otherwise be allocated to professionally more attractice activities. The quality control problem is a problem of professional mores.

THE QUALITY CONTROL PROBLEM: NEEDS VS. REQUIREMENTS

The EIA publishes projections of energy production, consumption and prices for future time periods: for next year, 20 years or more hence, and times in between. These projections propose to account for a host of relevant factors: geology, technology, international events and the nature of specific markets for energy products; the general performance of the domestic economy in terms of production, price levels and finance; stocks of energy producing and consuming capital, their location and changes in these; a variety of government initiatives, both those in place and those proposed.

These projections can support the Government's propensity to manage the energy system today in response to sudden developments such as embargoes, strikes and severe weather. They can help evaluate the consequences of legislative alternatives with impacts spread over many future years. They can inform private sector planning as it needs to take such future events into account. And, they can help evaluate the economic implications or desirability of technological change as it pertains to energy.

The usefulness of such projections seems self-evident; however, the projections prepared by the EIA and its predecessor organizations have never been entirely free of dispute. In part, the issues raised have been of the most normal and natural kind: a critical scrutiny of assumptions and methods. Any complex or at least detailed analysis can be carried out more than one way. Of the alternatives available there would usually be relative strengths and weaknesses among them. To select one alternative among several would necessarily entail some form of compromise and an opinion about the relative importance of the various attributes of the analysis problem. Since opinions will differ, the actual conduct of an analysis will always suffer in some respect in comparison to alternatives foregone. The continual review of alternatives and relative strengths of available methods is entirely proper; indeed, it is the hallmark of good professional practice.

As a matter of events, the Applied Analysis quality control program was not developed due to a pressure to establish a system of commonly agreed to professional review. Instead, the program has grown in response to a need for disclosure: the documentation of models and analyses and the establishment of a means for others to replicate analysis results.[4] Of course, disclosure in this sense is a prerequisite to professional review: so that there is no inconsistency. The point of these particular remarks is that disclosure itself is expensive, very expensive; perhaps on the order of 30 percent of the cost of developing a model and using it in an analysis. The need to budget for such nonproduct-related activities has been difficult to accommodate. To be clear, budgeting for disclosure occurs <u>prior</u> to budgeting for assessment: assessment costs perhaps another 30 percent, verification and access yet more (see the next section for definitions of these terms).[5]

The first paragraph of this section enumerates a number of important factors that (potentially) ought to be taken into account when projecting energy prices and quantities. How to do so? The analysis strategy chosen by Applied Analysis has been to construct an <u>integrated</u> model which takes the factors at issue into account <u>simultaneously</u>. This is difficult (i.e., expensive) to do. The result is large in the sense of embodying many numerical elements; the model is detailed (even complex) in the sense of embodying a host of mathematical structures that represent the attributes and rationale of energy market performance and the technologies of energy conversion and distribution; the model is naturally on a computer; the model is computationally ambitious in the sense that it may take a long time to solve relative to the total time available for an analysis (by "solve" I refer to the time--days and weeks--taken until the model is viewed as executing "properly" rather than the computing time taken to execute the model in a mechanical sense); and finally, the model solution is difficult to

interpret in the sense that the "answer" embodies many numerical elements (thousands upon thousands) which all vary with respect to the detail and complexity of the system.

The point here is that there is rather a lot to disclose.

In terms of needs it is agreed that the large, simultaneous approach is necessary in order to account for the host of factors that could potentially be important to an analysis. Further, it is self-evident that if an analysis capability is to be useful (e.g., sustained as an expense to the analysis process) then it needs to be responsive... results must be available in a form and at a time that will actually serve those seeking to make a decision. Alternatively, as an absolute minimum, it would seem that a requirement is that the nature of the analysis process must be disclosed. There remains a legitimate dispute as to what, exactly, is to be meant by disclosure; in what form and disclosed to whom? However, it is difficult to argue that the precise nature of an analysis should not be known to anybody.

These ideas should communicate a substantial problem. In principle, most agree that an analysis system and its use should be well documented. As a practical matter, such disclosure is very expensive, and from the standpoint of limited budgets, decisively competitive for the resources otherwise to be utilized in producing results. The needs and requirements of conducting good, demonstrable analysis are in conflict.

A TAXONOMY OF ISSUES

It is worth reviewing briefly what the issues to be resolved by quality control activities are, presenting some jargon with respect to these and, as a result, providing a frame of reference within which to place quality control activities to be discussed in the next section. Terms to be discussed are: documentation, verification, validation, the credibility of a model, access and computer aids for analysts.

Documentation -- this refers to written materials which describe a system in place (model documentation) and how a system is used (analysis documentation). Issues sometimes in dispute concern level of detail, intended audience and contents, such as also describing a model's rationale, precedents, and sources of data.

Verification -- the activity of determining if an analysis performed by the computer (or otherwise) is what was intended. If an analysis is to be verified, a separate statement of "what was intended" to be compared to a computer program prospectively or to some other statement describing what was (is to be) done is, seemingly, required. There is some dispute here since some argue that computer code is iteslf sufficient documentation.

Validation (viz: assessment or evaluation) -- the activity of determining the "correspondence" between a model and "reality." Many concepts can be at issue here. I find it useful to think of "validation" as the enterprise of determining how a model performs ... what is (and is not) taken into account and a model's strengths, weaknesses, and sensitivities.

Credibility of a Model -- given that a model is documented, verified and validated, I believe a further issue remains: what uses of a model so described are appropriate and what uses are not? I will use the term "credibility" in this sense.

Access -- a reference to the ability of a third party to reproduce analysis results.

Computer Aids for Analysts -- a variety of automated procedures designed to help analysts work with large systems. Techniques in place or under development include the ability to query solution files, express, work with and audit model structure, and form *a priori* expectations about solutions (e.g., automated "laugh tests").

THE EIA QUALITY CONTROL PROGRAM[6]

The EIA quality control program for analysis is intended to cover all of the issues in the taxonomy. The program is briefly described below.

Documentation (Models)

Models are to be documented in terms of three categories of information: Model Summary, Model Description, and Model Data Base Description. The "model summary" is a brief, several-page description in a unified format for all models which is periodically published by EIA. In addition to a summary description, information on such as the computing environment and official EIA contacts is also given. The "model description" is a statement of the model's structure, processes, rationale, and precedents prepared for a professional audience. The "model data base documentation" is a complete statement of the data upon which a model depends, sources, internal transformations (e.g., aggregation), statistical attributes, and naming conventions prepared for a professional audience. Generally, documentation is required for any other quality control activity. The central problem, particularly for models under development and in evolution, is the availability of *current* documentation. Given that documentation is expensive, the availability of current, complete model documentation continues to be a pressing problem.

Documentation (Analysis)

The concept of a "model" compared to all substantive activities at stake in producing analysis results is necessarily artificial in some respects. The use of a model will entail some form of

manipulation in order to represent assumptions or other special characteristics of the problem being addressed. Intuitively, model refers to the enduring features of an analysis capability (presumably automated) in contrast to all of the actions (whether or not automated) embodied in an "analysis." The dividing line between the model and these other activities should be taken as a convention rather than something sharp and self evident. EIA requires that each analysis be documented. Such analysis documentation calls for citations to the appropriate model documentation reports as well as a description of how the models were used and what else was done.

Verification

The verification of a model must in part refer to auditing computer code. For large systems, the state-of-the-art for this activity is in early development. In addition to a line-by-line audit of the entire code, possibilities include sampling the code, running test problems with known outcomes and testing the ability to detect "errors" purposefully placed in the model. Verification is conducted routinely by model operators, but as yet there is no uniform structure for procedures or protocol for reporting results. As it stands, the EIA program in model verification is just beginning.

Validation

The many strategies for evaluating a model fall under this category. Since knowing what a model is, is required in order to evaluate a model, model documentation is a prerequisite to model validation. A fair proportion of model validation efforts initiated by EIA have been frustrated to one degree or another by documentation problems. The results have been beneficial in terms of a more precise understanding of documentation problems; however, validation results have been delayed therefore.

Given the model validation results that were achieved, a further issue concerns the degree to which such results actually make a difference. In this usage "difference" refers to consequent changes in modeling practices and/or an enhanced ability to understand and communicate analysis results. In order to be successful, validation activities must be conducted in close collaboration with those responsible for producing analysis and be supported in practice by the management to which both the validation and analysis enterprise are jointly responsible. Since the goals of those involved are different, such collaboration and support are not automatic. A process of learning and working together is necessary. At EIA this process is still underway.

Credibility of a Model

As a logical matter, I have always thought that if everything about a model were known, there remained the residual issue of what, given all of this information, the model should be used for versus what not. If documentation, verification, and validation may be viewed as a process of disclosure, some systematic way to use this process should exist in order to reach conclusions about a model's range of application and attendant risks. EIA is beginning an inquiry into this issue. Not surprisingly, raising this issue had to await a tempo of progress in the areas of documentation and validation.

Access

Programatically, "access" refers to arrangements for transporting a model via computer programs on tape and suppporting written materials to others. The basic idea is to enable others to reproduce analysis results. A variety of purposes are served through such access. Published results can be double checked by those interested. The basis for results can be inspected in detail and alternative results can be derived from the same system for comparative purposes (comparisons of particular interest to a third party, but not of sufficient general interest to have been prepared by EIA). Finally, an analysis capability developed at public expense can be made available to the public. Although there are difficulties in access due to differences in computing environments, subject to these, the EIA access program is in place and successful.

Computer Aids for Analysts

In many ways the EIA has been a pioneer in using large, automated systems for policy analysis. It has developed that the use of large systems is constrained by their ability to be understood. Retrospectively this problem seems straight forward; however, it was not anticipated and there is very little state-of-the-art to draw upon for its resolution. Simply put, it is difficult to fully comprehend an analysis result with thousands of numerical elements based upon tens of thousands of lines of computer code representing a (bewildering) host of assumptions, parameters and mathematical structures. The EIA initiated pilot projects and solicited professional interest in developing the theory, and its means of application, in what has been termed "computer-assisted analysis and model simplification."[7] As appropriate to a new field results are, as yet, tentative and suggestive rather than decisive. Unhappily, this program has been especially sensitive to retrenchment due to budget constraints.

Other Quality Control Activities

The activities above stand as programs. The emphasis of their development has been that of initiating institutional change; that is, codifying, and from a budgetary standpoint defending a set of "quality control activities" which would routinely be part of the analysis process. To be entirely specific, the idea is that of developing activities which are agreed to in advance to be conducted for any analysis product and in relation to which there is (within reason) prior agreement about next actions given alternative outcomes of the quality control activities. For example, to agree that models are to be documented subject to certain guidelines, and if they are found not to be, to refrain from using them until they are.

An alternative is "professional review." Perhaps "analysis" is so special that its evaluation in terms of *a priori* criteria is not really sensible, or even possible. Instead, "quality control" might be viewed simply in terms of its proportion of the total analysis budget and in practice constitute the involvement of an appropriate group of (independent) specialists who would review an analysis effort (or a model proposed for use) and in the spirit of confronting specialized (nonrecurring) problems would reach tailored, specialized conclusions.

I do not subscribe to this approach in general; however, there are many advantages and certainly instances where such is proper. To begin with, reviews can be initiated immediately, and hence, results can be achieved quickly. Further, constructive interaction among all involved (reviewers, modelers, sponsors, ...) is not all that difficult to arrange. As a result, the outcome and implication of the review should be relatively understandable to those responsible for taking actions based upon the outcome of the review. Due to interaction with modelers and sponsors the reviewers would not tend to dwell upon issues whose resolution was essentially infeasible for one reason or another. Finally, all involved would be out of the (apparently unpleasant) business of deriving generic "standards," "guidelines," or "procedures" for analysis quality control.

These are very substantial strong points. The EIA and its predecessor organizations have sponsored numerous professional reviews which have been successful and of great benefit to the organization. Professional review as a program is now part of the clearance process for Model Documentation Reports.

Without arguing against these advantages, I will nevertheless claim that they are deceptive. I believe that quality control should be programmatic and structured in advance to the highest degree possible. I try to make this case in the section to follow.

QUALITY CONTROL FOR ANALYSIS

If an analysis result is published by a professional organization, what should it be taken to represent? What, just exactly, should be meant by the term "good professional practice?" For energy analysis the "profession" at issue is some blend of computer science, economics, mathematics, operations research, and statistics with rather explicit dependence upon geology, engineering and other "harder" sciences. The basic problem at issue is that of analyzing resource allocation through the mixed enterprise of a private ownership economy and government management. To a high degree, the techniques involved and the expertise required are generic and not particularly specific to energy compared to other allocation problems (energy as an exhaustible resource does call for differences in analytical approach compared to resources which are not exhaustible). Due to advances in computer technology, the ability to conduct detailed integrated analyses of market systems and their management is somewhat new. (I refer to micro- rather than macroeconomic systems.) It would seem that for this new application of (nonexperimental) quantitative methods there is little precedent for professional practice.

I believe that the generic analytical enterprise implemented for the case of energy as an exceptional effort will ultimately become a standard analysis technology to be utilized in many problem areas. The professional issue now unresolved is what as a professional matter should such a technology be taken to represent? What, precisely, should be taken to stand behind the results so derived? What do the results embody? Professional review as a quality control strategy, even in the ideal, will only serve to inform about a particular use of the analysis technology. Such review per se does not resolve the professional issue of what analysis results are supposed to represent.

Without resolution of this professional issue, it will be (and certainly has been) difficult to manage the results of quality control activities. In particular, "modeling" versus (such as) "validating" does not refer to different types of individuals, but rather, refers to different assignments made to the same professional cadre. Accordingly, a dispute between modelers and validators cannot be resolved decisively by management unless the issues at stake can be clearly related to an administrative structure agreed to in advance. An ample variety of issues exist, professional and otherwise, which will survive any administrative filter; it is these that should occupy the discretionary professional and managerial resources. As it stands, quality control issues are almost always exceptional and represent an excessive burden on managerial resources if they are consistently raised and are, likely as not, not raised at all.

The problem is, thus, the professions' concept of professionalism. I have every confidence that the basic intent of EIA's quality control program is in its main attributes beyond professional dispute. It is self-evident that models and analyses should be documented; that some systematic and demonstrable care should be taken to verify that the model on the computer is what is intended; that a model's precedents, rationale, sensitivities, and attributes important to its interpretation and use should be understood; that a model's use for a given purpose be in some sense demonstrably proper with consequent limitations well understood; and, if the resources are otherwise available, reproducibility through access is the hallmark of good science. It is difficult to imagine a graduate seminar at which these issues were somehow encountered spending very much time in their debate. They would surely be acknowledged and the proceedings move on to other topics. Yet, I am unaware of any professional group advocating that their (no doubt very good) analysis results be withheld due to a deficiency in any or all of these quality control issues. If the profession cannot take a stand on these issues, management cannot either.

I make my case on this last point if in no other way. Analysis results (produced as a best effort) are routinely presented for release although some or all of the following prevail:

- their basis may not (ever) be demonstrable to others or even recoverable from the analysis process that took place;

- what was actually done on the computer is not audited against what was intended (hard to do if there is no separate statement of intentions);

- the methods utilized are untried and associated strengths, weaknesses and sensitivities are essentially unknown; and

- there is no systematic determination by anyone that the methods used are the "best" available and in fact are sensibly useable for the purpose they were prepared.

I am as aware as anyone of the time and other resource pressures that constrain what analysts can do. If the profession will accept assignments with budgets that necessitate omissions such as the above, management will suppose that such practices are reputable. Quality control thus becomes a set of good ideas to be accommodated if time and money allow. This is unacceptable.

The solution to this problem requires collective rather than individual action. No professional directing an analysis effort

can preemptively suspend productivity due to a quality control requirement not generally practiced elsewhere. Given the constraints on analysis that sometimes are truely unavoidable, I do not question the actual merit of the results achieved. In many instances, quality control thus becomes the process of demonstrating merit. Without such a demonstration, good analysis results are often easily dismissed by those who wish to advocate conclusions rather than discover them. Of all the costs of insufficient quality control, perhaps this last is the most alarming: good advice going unheeded because its basis and nature are not disclosed.

The profession must take a stand on the manner in which we are to do our work if we are to do it at all. Such a stand was not taken at this symposium nor the last. We should do it soon. A next symposium should decisively resolve this issue at the very least.

NOTES

[1]The author is Director, Office of Analysis Oversight and Access, Office of Applied Analysis, Energy Information Administration (EIA), U.S. Department of Energy. The views expressed in this paper are the author's and do not necessarily reflect those of the U.S. government.

[2]See [1].

[3]I emphasize this point again: the views expressed in this paper receive their emphasis entirely due to my own personal experience. Notice for example that the Oak Ridge National Laboratory report, [2], cited in the first paragraph was sponsored by the Office of Energy Information Validation.

[4]See Public Law 94-385, "Energy Conservation and Production Act," Title I, Part A, Section 113, (August 1976). The point here is that critics of the analysis process did not maintain that the quality of the analysis was low; instead, they observed that an assessment of quality was not possible due to documentation problems.

[5]These percentages are rough estimates based upon experience and conversations of the author with those responsible for "code assessment" at the Idaho National Laboratory. The dividing line between "modeling" and "assessment" (i.e., "validation") is due to convention rather than inescapable logic. Some assessment projects have cost as much as developing the model at issue. As a rule of thumb, given current practice in terms of the stage at which a model is "ready for use," all related quality control measures probably should cost as much as bringing the model to this state of "readiness." This cost would be reduced as quality control becomes more routine and as certain "assessment" activities become part of the "model development" process.

[6]I emphasize again that I refer only to activities within EIA's Office of Applied Analysis.

[7]The EIA sponsored a symposium on these issues at the University of Colorado (Boulder) in March 1980. The Proceedings are scheduled for release in May 1981.

REFERENCES

[1] Lady, George M., "Model Assessment and Validation: Issues Structure and Energy Information Administration Program Goals," Validation and Assessment Issues of Energy Models, Saul I. Gass, Editor, NBS Special Publication 569, U.S. Department of Commerce, National Bureau of Standards, Washington, D.C., February 1980.

[2] Weisbin, C.R., et al., "An Approach to Evaluating Energy-Economy Models," ORNL-5742, Oak Ridge National Laboratory, Oak Ridge, Tennessee, March 1981.

VALIDATION AND ASSESSMENT PROCEDURES FOR ENERGY MODELS

Jerry A. Hausman
MIT

A few years ago when a colleague and I were discussing a recent book by a well known sociologist, my colleague somewhat derisively characterized his work with the phrase, "he is a person who writes books about books about books." My paper today has that aspect to it: I am considering the assessment and validation by groups of researchers of yet other groups of researchers' energy models. Perhaps a step back from direct scientific involvement adds perspective, so that I have tried to think about the important issues which arise in the assessment and validation of energy models. Although I am two times removed today from the real scene of the action, creation and testing of the models, I will draw on my previous experience being once removed as an assessor of energy models and also being at the scene of the action as a creator of energy models.

My main focus will be on that part of the EIA assessment and validation program which deals with the PIES/MFS model. [1] I must admit to my role as one of the initial assessors of PIES in my 1975 Bell Journal article.[2] I have maintained an ongoing interest in the progress of the PIES model. Thus, I feel that it is a good point in time after five additional years have elapsed to see what additional knowledge we have about PIES. Of course, PIES itself has undergone some significant changes in that period. Since I intend to limit my remarks to assessment and validation of the PIES model, the reports which I consider are:

(1) National Bureau of Standards (NBS), "An Annotated Restatement of the Methodological Description of the Midterm Oil and Gas Modelling System";

(2) Carl Harris and Associates (CHA): "A Sensitivity Analysis of the Domestic Oil and Gas Supply Forecasts Published in the Department of Energy's 1978 ARC";

(3) D. Freedman, T. Rothenberg, and R. Sutch (Berkeley): "Midterm Energy Demand: The RDFOR Model"; and

(4) MIT Energy Lab (MIT): "An Evaluation of the Coal and Electric Utilities (CEUM) Model" [3].

I found each of these reports to be of interest, and I feel that each report increased my knowledge about PIES. The NBS report is almost solely an effort in validation. I will consider the issues that it raises first. The other three reports have varying proportions of validation but are more concerned with assessment. I will consider questions of assessment of energy models at greater length since the scope of inquiry is considerably larger than in validation of energy models.

2. Energy model validation: A recurring confusion arises when validation procedures are discussed because of the meaning of the word "valid". Valid arises from the Latin word to be strong or powerful. Thus, valid is usually taken to mean able to withstand criticism or objection with accompanying legal connotations. But, we know that no model, even in physics or astronomy, is a perfect description of reality. Thus, every model depends on approximations. How good these approximations are and problems which may arise because

of the approximations belong to the realm of assessment procedures. If the EIA or other Government agencies attempt to validate energy models in the sense of deciding whether they are true descriptions then we know the outcome on a priori grounds. Even if they adopt seemingly more modest goals of deciding whether the models are "good enough" to withstand criticism or objection, I feel that the effect will become hopelessly entangled in the establishment of standards. Consider the degree of accuracy possible and even needed in models of population projections from HHS and models of satellite performance by NASA. Similarly wide deviations exist in energy models. Thus, to establish standards for validation which depend on the quality of the model is misguided. Quality decisions belong to the realm of model assessment where sets of standards would be much more flexible. We need to establish much more narrow and rigorous standards for model validation procedures.

Thus, I propose that model validation efforts concentrate on two main areas: (1) data, (2) the correspondence of the logical and mathematical structure as described by the model's creators with the actual computer code which implements the model. Here we would look only at the validity of the model as designed and described in its documentation.[4] Questions of the appropriateness of the logical and mathematical structures, statistical methods, judgements of model uncertainty, and the appropriateness of the data would all enter the assessment program. In validation procedures we would consider model performance conditional on acceptance of model structure. Within the restricted confines of such an analysis, I will propose some possible cases for performance standards.

A. Data Validation: Almost all models (except perhaps for systems dynamics models) are based on data. Many energy models have much greater data requirements than the average model because of their attempts at fine geographical resolution. Thus, questions of data validity become quite important. First, validation procedures should undertake obvious tasks of source validation. Data items from Government agency sources need to be checked. Changes in definition of data elements over time and procedures to account for them need to be validated. Units of account also need to be checked, e.g., constant dollars or current dollars for a price series. Note that this requirement leads to certain standards in documentation which have not been met in all areas of the PIES model. Next exogenous parameters of the model, sometimes referred to as engineering constants, need to be checked. Procedures might also be developed to do computer checking of data in order to catch certain gross errors. These errors seem to enter data sets in transcriptions and are usually, but not always, discovered in the model building process.

These procedures are relatively uncontroversial. My next suggestion on data validation may be less well received. Much data used in energy models does not have a truly precise definition. For instance, "the price" of gasoline in a given area for a given time period does not exist for reasons of geographical and temporal variability. Thus, even on a theoretical level index number problems arise. Various weighted averages are used as index numbers where the theoretically correct weight function usually depends on the particular demand or supply functional form which is being estimated.[5] The description of the data should contain a mathematical or descriptive statement, e.g., Divisia index, so that the index formula which has been used can be checked. Part of the validation process would involve checking for the correct use of the mathematical formulae of the various index numbers. Now I reach the somewhat controversial aspect of data

validation. A significant proportion of the desired data for a given energy model may not exist. Synthetic procedures, usually based on regression analysis, are often used to create the data which does not exist. Thus, the price of a particular fuel in Region 2 for the years 1960-1965 say may be the conditional expectation of the fuel price from a regression specification estimated over the years 1965-1975 when the data does exist. Questions of the appropriateness of such data creation techniques usually have no standard answer. I suggest we leave a discussion of questions of appropriateness to the model assessment procedure phase of evaluation. However, the model validation procedures should check on how all such synthetic data were created. They should further insist that the documentation of the model explicitly state, in mathematical form, all procedures used to create the synthetic data. A possible user of the model data would then know explicitly that he does not have the price of fuel; but instead, he has a synthetic data element created by a procedure included in the documentation. Caveat emptor. The standard which should be aimed at would permit a reconstruction of all data elements by use of the formulae contained in the documentation on the original sources. Then, even though the user really does not have the 1963 price of jet fuel in Region 2, he will have access to the formulae for what he does in fact have from the data used in the model creation.

B. Model Validation: Again, I believe we need to eschew questions of model appropriateness if we hope to establish standards for validation. What I feel that model validation should concentrate on is whether the computer code does what the documentation claims it does. In building large models it is inevitable that mistakes enter the code. Especially with a model as large and complex as PIES where many revisions and patches have been made over the years, it is quite important that users be able to assume that model output is valid within the confines of the model structure. It is quite interesting that three of the reports which I reviewed, the NBS study, the CHA study, and the MIT study all discovered mistakes in the implementation of the model in the computer code. I do not believe that the computer code for an energy model can be taken as valid in the sense of being the statement of the model builders intentions. Internal contradictions within a model are a certain sign of mistakes, e.g., use of current dollar prices in one section of a model and constant dollar prices in another section. Many other types of problems are found when a model's computer code is scrutinized by outside reviewers. For instance, if the model documentation claims to enforce a certain type of regulation on energy markets, the validation procedure would check the computer code directly to see if the regulations have been implemented in all parts of the model correctly. As a model builder I have found this type of validation review to be extremely valuable. Too often under the pressure of deadlines, mistakes enter models but are not discovered by the model creators. I foresee an important role for validation procedures which test the computer code for internal consistency and absence from mistakes.

To briefly review the four reports on PIES that I listed previously, I found the MIT Energy Lab evaluation of the CEUM model to be best in model validation. However, it did not pay particular attention to data validation questions since it focused more on assessment procedures which I will discuss in the next section. I found the NBS study of the MOGMS study, which is closest to a pure validation exercise, to be somewhat disappointing. So far as I can see, except for a single exception, the NBS study did not turn up much new about the operation of the model from how it has been described in

EIA documentation. Almost no data validation was undertaken, and the model validation is too limited in size to be a conclusive validation study of the MOGMS model.

3. Energy Model Assessment: I have tried to keep the scope of validation procedures narrowly defined. The reason for the restricted scope is the idea that standards could be established for validation. Government agencies could then judge models as valid. But, I think a more important exercise, which I want to keep distinct from validation, is model assessment. Here all the questions of model adequacy would be discussed. Even if a model is technically correct in its execution, we might still not find it to be adequate due to data, model assumptions, or model structure. Model assessment procedures often lead to model improvements when the model builders receive suggestions of better ways to look at problems. But since questions of model adequacy are based primarily on value judgements, I believe it is unlikely that the Government could set standards for certification of model quality. In fact, some models are adequate for a certain use but become inadequate when used for problems that the model builders did not intend them for. Thus, I will suggest a set of areas that model assessments might well consider. But in no sense do I put them forward as standards. Some excellent model assessments can be quite broad in scope like the MIT assessment of the CEUM model or more narrow in scope like the CHA assessment of the MOGMS model. I feel both reports contributed significantly new information to our understanding of the respective models.

A. Data Assessment: I begin again with data; but I now consider data assessment rather than data validity. In the review process for the READ model which I took part in for EIA last year, the data might have been judged valid. However, it seemed to the review panel to be grossly inadequate. In fact, the READ review led to one of my favorite exam questions of recent years. By creating synthetic data the model creators had supposedly discovered a superior estimation method to least squares estimation, even when the Gauss-Markov assumptions held true. But the READ case was almost too easy since the attempted level of geographical resolution was counties. More difficult cases arise in judging data adequacy within PIES. For instance, the Berkeley group judges the data for the RDFOR demand model of PIES to be inadequate. I agree with their judgement in many respects. Yet, I feel that they are unduly harsh in other respects. For instance, if we consider gasoline demand models at almost any level of aggregation, problems arise with the data. Suppose we have accurate quantity data by state, which is reasonable since tax revenues are collected on all gasoline sold for motor vehicle use.[6] No matter what type of price data we attempt to use, problems of aggregation will arise. For less aggregation, problems of quantity data arise while price aggregation problems worsen as we move to larger geographical regions. Nor do I see how the problem will be resolved unless some misguided DOE regulator finally suceeds in forcing a uniform price of gasoline in the entire U.S. But my judgement is that we could still estimate a useful gasoline demand relationship so long as the price data has been consistently constructed over the time span of our model. Questions of data adequacy will always be present, but I do not feel that a nihilistic approach to energy models is proper. Overall I do not feel that the PIES model has attempted excessive geographical resolution in the demand and supply sub-models. Many of the problems which exist in PIES would not be

alleviated by going to a model with less resolution, while other new problems would certainly occur.

B. Economic Structure: I use the term economic structure to encompass all the methodology used in a particular energy model to depict the workings of the various markets. I feel that evaluation of the economic structure of a model is the most important part of an assessment study. Models rely on simplifying assumptions. Assessors need to decide if the model assumptions are sensible, and if the model structure is adequate to study the problems it is designed for. For instance, I would like to list a few of the questionable assumptions of the original PIES model which I identified in my Bell Journal article and consider their possible importance with 5 years of hindsight:

(1) The energy model is not integrated within a macroeconomic model. Subsequent events have shown that the raising of the oil price can have very large effects on the macro economy through inflation induced by rising energy prices and subsequent government policy measures. In fact, some economists attach an extremely high shadow price to marginal barrels of oil, on the order of $75, because of the inflationary consequences of repeated OPEC price escalations. These macro effects then feed back into demand for energy via the level of economic activity. Since 1980 will probably see an acute recession, I doubt that the structure of PIES can adequately capture what will happen to energy demand this year.

(2) Electric power production through increased coal and nuclear generation was assumed to play an important role because environmental and regulatory constraints were not assumed to be especially binding. These assumptions were remarkable because even in 1974 when the model was constructed they were having an important effect. In fact, these constraints have become even more binding, bringing nuclear construction to a virtual halt. Also, use of coal burning plants has not proceeded to the degree expected by the model.

(3) The oil and natural gas supply models used average rather than marginal decisions and basically set nonassociated natural gas supply to be totally price inelastic because of exogenously assumed target rates. As it has turned out, large amounts of new exploration and drilling for both oil and nonassociated natural gas have occured in response to the higher prices.

My point of this brief review is to demonstrate that certain key assumptions are usually more important than exact details of data quality, statistical methods, and other factors which receive the majority of attention in assessment studies. I think that assessment of the adequacy of these assumptions is vital to a good understanding and judgement of model quality. Unfortunately, many key assumptions cannot really be tested within the confines of the model since they form defining aspects of the model structure. Still, to have an informed judgement on how good the models are, we need to examine their economic structure.

The MIT Energy Lab report does an excellent job of evaluating the economic structure of the CEUM model. I was somewhat disappointed to find how little attention economic structure received in the other reports. The CHA report explicitly stated that it did not feel this type of evaluation to be within its scope of work. I would like to encourage EIA to sponsor more evaluation in this area. The results are likely to be less clear cut than more narrow technical evaluations. Still, I feel that our knowledge of model capability and quality may be increased most by just such studies which evaluate the structure and assumptions of the energy models.

C. Statistical and Computational Methods: These methods are more narrowly technical than the previous two categories, but still important questions can arise. I will concentrate on statistical methods since I have more knowledge about their application than I do about computational methods. The Berkeley report on the RDFOR model makes numerous criticisms about the techniques used to estimate the model. They claim that the model has vastly too many parameters and that simultaneous equation estimation techniques should have been used rather than least squares methods. The first criticism is a matter of judgement, and the authors might have done more to buttress their claim than count unknown parameters. The second criticism could be quite important since the identification problem only allows us to separate a supply from a demand curve using simultaneous equation estimation techniques. But again no evidence is provided to allow us to judge how serious this problem may be. Other criticisms of the Berkeley report such as incorrect stochastic specification may also raise problmes with RDFOR. But in this area I feel that a good assessment report should be expected to reestimate the models to see how important their criticisms turn out to be.[7] In the areas of statistical and computational methods, the data exist to try out different ideas. By attempting different methods of estimation or different stochastic specifications, the change in model coefficients and forecasting ability can be assessed directly. Also, formal statistical tests can be performed to decide whether the differences attain statistical significance for a given size of test.

D. Sensitivity Analysis: I would immediately like to commend the CHA analysis of the MOGMS system. I feel they have conducted a well designed assessment of an energy model which would serve as a good starting point for further efforts. The intelligent use of Monte Carlo simulation techniques and response curve analysis indicated how some key uncertainties can be resolved and important parameter assumptions can be found through sensitivity analysis.[8] Too often sensitivity analysis appears to be a random walk type procedure in first one set of parameters is changed, then a second set of parameters is changed, and so on without an over-riding analytical approach to the problem. A well organized attack on the problem seems to offer a much higher probability of success. For instance, one of the key findings of the CHA report is the importance of the economic parameters assumed in the model. To date considerable attention has been focussed on the geological uncertainty while too little attention has been given to the economic situation. To the best of my knowledge no oil or gas supply model accounts correctly for the economics of drilling and exploration. The CHA results indicate that research effort in this direction might be highly rewarded. Such a finding is what we hope for in a well conducted assessment study.

Thus, in the area of sensitivity analysis I recommend adoption of a more structured approach than many studies have taken. The cost of sensitivity studies can be decreased by remarkable amounts if well designed Monte Carlo techniques are used. Also, response surface analysis is an important tool which should be used more often. Both techniques are widely used in statistics and econometrics. They have seen an increasingly wide use in econometrics where certain mathematically intractable problems have been well analyzed with Monte Carlo and response surface techniques. Many of the problems bear a close resemblance to problems which arise in sensitivity analysis of energy models.

E. Benchmark on Real Data: Perhaps the most astounding deficiency of assessment studies is the almost total neglect of what is happening in the real world. This criticism is extremely important for PIES. The original model was run for 1977, 1980, and 1985. We have passed 1977 and in 14 months time will have 1980 data. How well PIES can do using available data up to 1975? I would permit alteration of assumptions within the current PIES framework and not hold it to the 1974 assumptions. But where would the model go seriously wrong for a real year in the future? I feel we could learn an enormous amount from just such an exercise. I strongly recommend that EIA sponsor this type of assessment.

Otherwise, I cannot see valid alternative methods to assess how good energy models are. Each quarter the macro models are put to the test. Yet, in energy models much research is comprised of running different models on the same energy market far off into the future. I see little use in such an exercise. It may well be the model which gives outlying forecasts which is the most valuable. If it turns out to be more nearly correct, the required changes will be all that much greater. Running three models together and finding that two are close together and one far apart tells me little information except that I should examine the model structure closely to identify the reasons for the differences. The exercise is quite unlikely to convince me that the outlier model is less good than the other two models.

I would like to end on this point. Because of lack of data, many energy models have resorted to techniques of data synthesis. Thus, some of their variables are unobservable independent data series. But at some level of aggregation they should be producing forecasts of available data. Otherwise, they are rather uninteresting models. I see a crucial neglect in analyzing how well the models do perform in reality. I need some information here to decide if, in fact, energy models are credible.

FOOTNOTES

[1]My chosen acronym for the overall model will be PIES which I know is officially outdated but has that advantage of easy recognition from the plethora of other model designations.

[2]Jerry Hausman, "Project Independence Report: An Appraisal of U.S. Energy Needs up to 1985", Bell Journal of Economics, 6, 1975.

[3]I am associated with the MIT Energy Labortaory, but I have not been involved in their energy model review program since 1975.

[4]I realize that a problem may arise because documentation is often not done by model builders. But, I feel that model builders should be held responsible for the accuracy of the documentation. Thus, the description of the model, either verbally or mathematically, should be the primary document to consult.

[5]A recent exploration of these issues is contained in E. Diewert, "Exact and Superlative Index Numbers", Journal of Econometrics, 8, 1978.

[6]Even here problems arise because taxes are not levied on gasoline for recreational use like motor boats. But some compromises are necessary.

[7]The Berkeley authors promise such a study in the next phase of their asessment.

[8]I do have two minor quibbles. First the method of antithetical variables did not seem to be used which can further reduce computer time. Second, the use of (-1,0,+1) for functional changes in the response model seems incorrect. Two separate variables corresponding to the -1 and +1 cases seems a preferable specification.

ENERGY MODEL ASSESSMENT PROCEDURE DEVELOPMENT PROJECT

Richard H. F. Jackson
National Bureau of Standards

The primary concern of this project is the development of methodologies for performing model assessments. This project fits into the overall mission of the National Bureau of Standards as we are concerned with the development of procedures, guidelines, or standards for doing such assessments. It also fits into the overall mission of George Lady's office at the Department of Energy. That office has the responsibility for overseeing the process of model building and model using. To that end, DOE has sponsored NBS assessments using specific DOE models. The first model to be assessed was the Midterm Oil and Gas Supply Modeling System. The results were described in a series of 11 reports (to be reviewed below), and were discussed in the two reports that Drs. Hausman, Richels, and Shanker reviewed earlier.

Before I go on, I would like to thank the organizers of this conference for this opportunity to comment on the comments made by the reviewers on the comments we made on the models in use at the Department of Energy. Also, I would like to thank the reviewers for the comments they made. This is the first opportunity for feedback on our work, and, of course, we welcome it. Unfortunately, I must spend some time clearing up some misconceptions about our work.

It seems there was some confusion about the Analysis Quality Report that NBS submitted. To begin, it is not true that the report entitled "Sensitivity Analysis of MOGSM Results" prepared by Carl Harris Associates (CHA) and the report entitled "An Annotated Restatement of the Methodology of the Midterm Oil and Gas Supply Model" were separate AQR's. The CHA report (as it has been referenced in this meeting) was prepared under contract to NBS and serves as the central paper from among a set of papers submitted as the AQR from NBS on the Midterm Oil and Gas Supply Modeling System. The Annotated Restatement is one of a number of other reports that are intended as appendices, or backups if you will, to the CHA report.

There is another report, which was not submitted, entitled "Investigations Into the Underlying Data for the Midterm Oil and Gas Supply Modeling System." This was referenced in the CHA report in which the results are given of an intensive investigation of the data supporting MOGSM. For our project, this served as the "data validation" effort that Jerry Hausman said should be done. Another report, "Data Extrapolation and Statistical Forecasting," contains the results of our investigation of the statistical estimation procedures used.

The backup reports review the conceptual formulation of the model and its assumptions, and the statistical estimation techniques used. They contain information that served as input into the experimental design used to performing sensitivity analysis of the outputs of the Midterm Oil and Gas Supply Model.

Only the two reports, the "Sensitivity Analysis" and the "Restatement of the Methodology," were submitted for review as it was unclear how many of the reports I just mentioned would be useful to the reviewers. Since the purpose of the AQR's was to assist George Lady's office in identifying the degree of confidence one can have in the numbers published in the Annual Report to Congress, I felt the reviewers would be more interested in our efforts to discuss those numbers and that confidence, and less interested in reviewing our efforts to assess the model and develop a methodology for model assessment. My intent, therefore, was for the reviewers to see the central report, which presents our sensitivity analysis of those numbers, and, if they were interested, to provide copies of the backup reports later.

I would like now to respond to a few of the specific comments made by the reviewers. Dr. Hausman felt that a serious omission in the NBS work was the lack of a data validation effort. I hope my comments above clarify this matter. He also reviewed the Annotated Restatement as a model validation report and discussed ways to improve it. I consider this a compliment because the report, as I mentioned above, was not intended to be a model validation report. Our goal in writing it was to discuss the methodology more completely than it was discussed in the documentation provided us, and also to discuss some of the assumptions made in the model. This report was written because such a description did not exist and we needed it to ensure that we understood the model as it existed, not as it was documented years earlier. We studied the documentation and held a series of meetings with the staff of DOE and the contractors responsible for building the Midterm Model. Our report began as an internal working paper and continued to grow in size and stature to the point that we felt it was a worthwhile addition to the documentation of this model. We thus chose to publish it.

Dr. Hausman also commented that he felt validation should be the act of comparing model documentation with computer code. I do not think the term "validation" should be restricted in this way. I would perhaps call that verification, but I don't want here to get into a discussion of taxonomy. I do, however, believe strongly that such an effort (whatever we call it) is necessary. It is unfortunate that we were unable to do more of it, but it is very difficult to compare code with documentation when documentation does not exist.

We began our project believing that assessors should not be required to read code in order to understand what a model is. The models in question change much faster than the documentation, making code reading mandatory. It is perhaps boring, pains-taking, and time consuming, but we discovered a number of simple coding errors in the process, thus improving the model.

Dr. Richels commented that the Analysis Quality Reports are too technical to be used by the policy makers for whom they are being written. I cannot comment on the technical background of policy makers, I have yet to be introduced to one.

At NBS our efforts to develop methodologies and to produce AQR's have been limited to techniques for model assessment. We believe that this is a necessary first step in the development of approaches for presenting the results of such model assessments. Perhaps these can be done pari passu, but we have not concentrated on this problem. We have conducted workshops on the meaning of model confidence. We have conducted workshops on model documentation. But we have not spent time with the policy makers for whom these reports are written in order to learn what their needs and desires are. Perhaps when we have learned more about how to conduct model assessments we will be in a position to condense the results into a format suitable for presentation to policy makers.

Dr. Richels also felt unsure about the value of efforts to make models portable. He said that model portability was perhaps an unattainable goal, since sponsors rely so much on the original developers of the model for their proper use, application, and interpretation. I agree with him that this is a difficult task; the state-of-the-art is such that models are not stand-alone entities. We have, for example, a report coming out describing our efforts to convert the Midterm Oil and Gas Supply Model from the DOE computer system to the NBS computer system; two very different machines. It is not an easy task even to perform the physical act of converting from one machine to another. Also, we have not addressed the question of the intelligent use of that model when it is separated from the original model developers. Nevertheless, I feel strongly that we must continue to require that models be made portable. In fact, this ties in with comments I had on another of Dr. Richels' points. He said that in his view policy analysis is more art than science. I feel that there are components, certainly, of policy analysis that are more art than science. But I feel that there are many other components that are very scientific in nature. While perhaps we might not be able to subject the artful aspects of policy analysis to peer review and scientific experimentation, we can certainly subject those other scientific aspects to the same rigorous requirements of scientific reproducibility and investigations that we all were trained to respect. Viewed in this light, model portability is an absolute necessity. That which is to be subjected to peer review must be capable of being viewed, reviewed, and reproduced by outsiders.

Large scale mathematical models are currently being used to help set national policy. They are here to stay. I think it is no longer possible to hide behind claims of "art," "too complicated," and "no time." We must recognize that these are scientific tools and techniques that must be subjected to the rigorous requirements of scientific reproducibility and experimentation, and so conduct ourselves as the professionals we claim to be.

COMMENTS ON VALIDATION AND ASSESSMENT PROCEDURES

Allen L. Soyster
VPI & SU

I have had some first-hand experience with many of the things that have been discussed here today. These experiences are not only with energy models but with some other models which I will mention.

To begin, I would like to make a couple general comments, and a couple of specific ones related to some issues that have surfaced here.

This morning, Ken Hoffman suggested that it was difficult to be both the performing artist, a member of the audience and a critic and that these tasks may well be separated and maybe should be separated.

I feel differently. I feel that to be a model developer and a model builder that one has to do all three of these things.

When one develops a model one doesn't develop the model for six months and then attempt to verify or validate it the following three months. I think it is an on-going back and forth procedure and there is a continual validation.

Just a short time ago at Virginia Tech we have been involved with the building of a non-fuels mineral model of the world copper market. It has nothing to do specifically with energy but the same issues relative to validation processes are applicable. Recently the Bureau of Mines' forecast U.S. copper production will be 3.2 million tons by the year 2000. We are currently producing at about a rate of 1.6 million tons per year.

In support of this non-fuel minerals model we have collected data on every individual mine, smelter and refinery in the U.S. plus information on other countries' supply of copper -- Zambia, Zaire, Chile, Peru -- a number of them.

Our own forecast developed from the non-fuel minerals model, developed much like the PIES model, is that there is no way under current technology that the U.S. can economically produce more than about 2.4 million tons of copper.

The 2.4 million ton estimate was not arrived at overnight. This number has been under scrutiny for well over a year. In terms of the model development, the verification was a day-by-day thing and I never really thought that at any particular time we were doing validation versus model development. They happen simultaneously. Preliminary results are followed by model modifications which are followed by further tests and so on. Hence, I feel that a model user should take some solace in the sense that in many cases model validation is an inherent part of the model building task.

It was also pointed out this morning that Saul Gass and Lambert Joel had a four-fold evaluation of model assessment. And, as I recall, it was documentation, verification, validation, and usability. The first, second and fourth, it seems to me, can be done in a fairly straightforward way. It is just a matter of resources and time.

The third item, validation, is the one that is the central issue here. For example, consider the notion of validation in terms of the Virginia Tech copper model. Although we are not really sure that the forecast of 2.4 million tons of U.S. production in the year 2000 is a realistic estimate, it is internally valid; the 2.4 million ton estimate can be justified and explained very well. In terms of costs, efficiencies, reserves and foreign imports the 2.4 million tons is consistent with current technology.

Permit me to return to another comment. Although I have not been involved with these quality assessment reports, I have been involved with the assessment activities within EIA which deal with historical data.

In particular, one of the things that Fred Murphy and I have done was to take one subpart of the PIES model, the electric utility submodel, and subject it to 1977 data.

Now there were no capacity expansion issues involved here. The question was---given this 10-region system, and given fuel prices that existed in 1977 and given load duration curves associated each of these 10 regions, does this submodel reasonably replicate the dispatch of equipment types with what happened in 1977?

This historical test was run for each of the 10 geographical regions of PIES which comprise the U.S. However, I will present only the national totals. For the nation as a whole, coal represented 46.5% of electric energy generation in 1977. When the model was run with 1977 data the percent generation by coal was 48.4%. The comparison by major fuel types is

	Actual 1977 Generation	Model with 1977 Data
Coal	46.5%	48.4%
Oil	16.9	12.0
Gas	14.4	14.0
Nuclear	11.8	12.2
Hydro	10.4	13.5

I should mention that 1977 was an extremely dry year in the Pacific Northwest. In fact, hydro generation in 1977 was only 75% of the average of the previous three years. If these numbers would have been adjusted to a normal hydro year and the remaining hydro generation proportionately spread over the other fuels, then I thought the historical data matched the model reasonably well.

ARE ENERGY MODELS CREDIBLE?

by

David Freedman[1]
Statistics Department
University of California, Berkeley

1. Introduction

In this paper, I argue that the forecasts from current energy models cannot be relied upon. Indeed, there is little hard evidence to show that such models work. And on *a priori* grounds, skepticism seems justified. The quality of the data inputs is often poor or uncertain, and in some cases the internal logic of the models is open to question.

Before developing these points, I have some comments to make about the institutional setting. A major part of the analytical effort at the EIA (Energy Information Administration) is devoted to making forecasts for the Annual Report to Congress. These are made for the short term (a year or two ahead), the midterm (five to twenty years ahead), and the long term (the year 2000 and beyond). These forecasts are produced with the help of very large mathematical models, for instance, the Midterm Energy Forecasting System.[2] A single forecasting run with this system takes several hours of computer time.

EIA models require very detailed data on energy production and consumption. In fact, their demand for data far outstrips the supply, despite the extensive data collection efforts at EIA and other governmental statistical agencies. (Of course, most of these data collection efforts were not designed to support modelling.) As a result, much data used by EIA models are synthetic, meaning that most of the numbers result from imputation rather than measurement.

Despite their complexity, EIA models are often unable to answer the policy questions of current interest. This is partly because the questions evolve faster than the models. Furthermore, if left to their own devices the models often produce unreasonable forecasts. To answer relevant policy

---

[1] This paper was written for the DOE/NBS symposium on validation and assessment of energy models, May 19-21, 1980. It grows out of an extended collaboration with Thomas Rothenberg and Richard Sutch of the Economics Department, University of California, Berkeley, on a project to assess RDFOR. This project was funded by the Office of Analysis Oversight and Access, EIA. Sections 3 and 5 of the present paper are abridged from project technical reports.

[2] This is a successor to PIES (Project Independence Evaluation System).

questions, and to keep forecasts on track, EIA analysts have to make numerous subjective adjustments to model inputs and outputs. Perhaps for reasons already given, EIA models are in a continuous state of development. There are always new questions that must be answered: the pressure is for more detail in the forecasts. Periodically the models must be reworked to give more sensible forecasts: an unreasonable forecast is usually construed to result from the model's failure to capture some detail of the energy market. EIA models start big, and they are growing.

What are the consequences for the credibility of EIA forecasts? To begin with, the accuracy of synthetic data is usually difficult to assess. The standard statistical procedures, like regression analysis, were designed for use with real data--and may behave differently when run on synthetic data. In particular, degrees of freedom and standard errors become very difficult to interpret. Worse, errors in variables can lead to serious bias in coefficient estimates--without any way to estimate this bias. This complicates the task of assessing models built on synthetic data.

Next, there is a tension between the idea of modelling and the idea of subjective adjustments. I do not for a moment suggest that EIA should be publishing silly forecasts just because they come out of a model. But this inconsistency in EIA's modelling philosophy is worth thinking about. And there is some impact on the credibility of the forecasts, because it is hard to assess the quality of judgment involved in changing intercepts between model runs, or splicing models together.

Another consequence of EIA's modelling philosophy is a shortfall in quality control work, and this too is quite damaging to the credibility of the forecasts. What kind of track record do the models have in forecasting? What are the statistical properties of the data? the equations? the fitting procedures? How do the models respond in simulation studies, or sensitivity analysis? Before they trust the forecasts, analysts outside the EIA will want answers to such questions. A start on sensitivity analysis was made in the 1978 Annual Report, and this is an encouraging sign. Varying the assumptions in the scenarios, in the forecasts of key exogenous variables, is an important thing to do. But as I will show later, EIA models embody many technical assumptions about the energy market. A lot of hard work needs to be done, to see how the forecasts depend on such assumptions.

A final point on quality control. Proper documentation is crucial in establishing the credibility of any analytical enterprise. For example, it seems possible to recreate the 1970 Census of Population from the printed record. Furthermore, census publications openly acknowledge the problems with the data; they report serious efforts at quality control, and at quantification of errors. This can be quite disarming, for it demonstrates that the problems have been considered. By way of comparison, it seems impossible to reconstruct the forecasting procedure used in the 1978 Annual Report from the accompanying documentation. EIA is in the awkward position of using models to make forecasts, but being unable to say just what those models are.

In large part, I ascribe the shortfall in quality control work and in documentation to the way the modelling enterprise is managed. EIA analysts are required to produce forecasts in great detail, and with considerable frequency. They are required to use very complicated technology in making the forecasts. They are required to elaborate that technology from year to year. Not unnaturally, quality control and documentation come to be seen as dispensable. So do the statistical proprieties. The nitty-gritty is getting the numbers out. This seems curiously impractical. The numbers do get out, but nobody quite knows what they mean, or how far to trust them.

2. The case against the models

In the previous section, I suggested that energy models lack credibility because they run on synthetic data, and because they are not documented at all well. I must also tell you that at least in some cases, the models themselves are seriously flawed. Even with good data and good documentation, such models could not compel conviction. This opinion is based on a careful assessment of a very small sample of energy models, and not a random sample either. As a statistician, I want to be cautious about extrapolating from such a sample. However, it may be useful to draw some general--if tentative --conclusions. To keep the discussion in focus, I will concentrate on econometric demand models. There are some major sources of uncertainty about such models. These are well known, but it may be useful to review them here, as we consider the strategy of using models in forecasting.

First, there are major events which influence energy markets but which are beyond the ken of econometric demand models. Recent examples include the Arab oil embargo and the Iranian revolution. Unforeseen events of this magnitude are apt to occur in the future, and are likely to throw the forecasts well off the track. This point is obvious, and the inside front cover of volume III of the 1978 Annual Report insists on it. Obviousness does not detract from importance.[1]

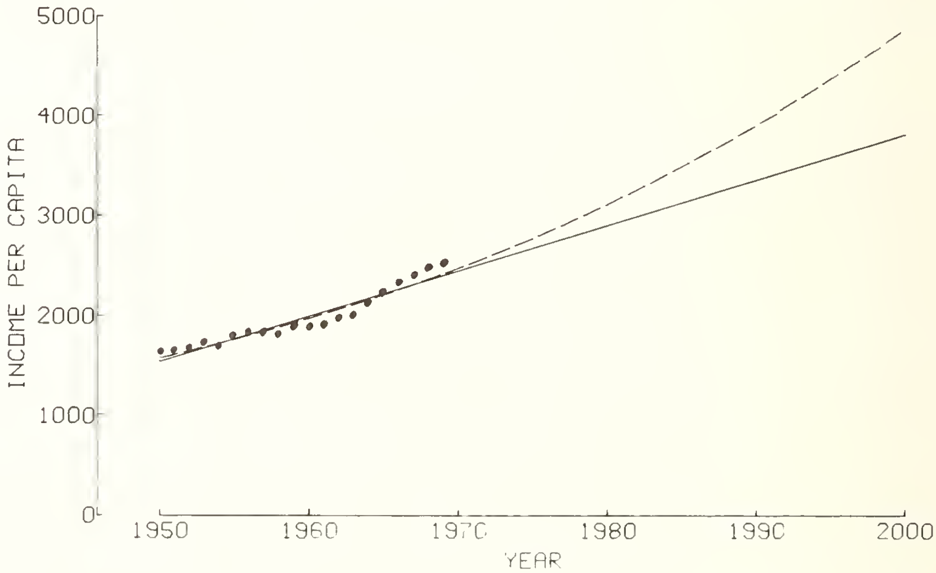
A second source of uncertainty about econometric demand models is that they have quite weak theoretical underpinnings. This is because economic

[1]As an interesting sidelight, I suggest that it is impossible to make useful estimates of the variance caused by rare events, because the fitting periods of the models are too short. For instance, let us make the assumption--which is quite favorable to the possibility of estimating variance by statistical methods--that rare events of a certain type follow a Poisson process, with a rate which is unknown and to be estimated. To get some idea of the orders of magnitude involved, supposed that in fact, events of this type occur at an average rate of one every ten years. In a fitting period of ten years, we have a 37% chance of seeing no such events, and estimating the rate of occurrence as zero--not a useful estimate. There is a 37% chance of seeing exactly one such event, and getting the estimate right. Finally, there is a 26% chance of seeing two or more such events, and hence being off by a factor of two or more in our estimate.

theory does not dictate specific functional forms for behavioral equations, although at any given time some forms are more fashionable than others. Likewise, the theory does not describe any very specific chance mechanisms for generating the stochastic disturbance terms. As a result, the assumptions imposed on these terms are necessarily somewhat *ad hoc*.

When fitting a curve to data, the functional form may not matter too much--if you stay where the data are. When extrapolating from the data, the choice of functional form becomes a critical activity. Usually, there will be several plausible specifications, and no way to tell either from the theory or the data which is the right one to use. But the different specifications will lead to very different forecasts. This is illustrated in figure 1 below, which shows the trend of *per capita* income against time.[1] Two different trends have been fitted: linear and exponential. Both track the historical data quite well, agreeing very closely during the fitting period (1959-69). By the year 2000, however, the exponential curve is 25% above the straight line.

Figure 1. In extrapolations, functional form matters.



This sort of point is important for econometric demand models, where somewhat arbitrary functional forms are used to extrapolate from a period of energy abundance to a period of energy scarcity. Of course, the example may seem quite simple-minded, by contrast with the apparent sophistication of an

[1] Disposable income in 1958 dollars. The data source is the *Economic Report of the President* (1974).

econometric model. However, putting in more variables and more equations will not help matters, when each equation just represents an additional postulate. In fact, multiplying assumptions may make matters distinctly worse, because sorting out the consequences of the specification becomes very difficult.

The extrapolation problem is crucial too, when a model is used to estimate the impact of a proposed policy change. Indeed, the policy initiative is apt to change the rules of the game, pushing the model into completely new territory. For this sort of reason, statisticians are quite suspicious about estimating the effects of interventions, on the basis of curve-fitting to observational data.

Of course, most of the time the world changes quite slowly. Regression analysis and the computer are powerful tools for discovering relationships in data, especially when there are few constraints about which variables to put in the equations, how to transform them, or what kind of lag structures can be tried. As a result, newly-fitted models are apt to track very well, for a while. And then they are apt to stop working, because the regression equations turn out to be artifacts of data processing, rather than laws of nature. (See Appendix A.)

Let me now summarize. At the EIA, generating forecasts with big models consumes a lot of resources: little is left over for proper quality control or documentation. Furthermore, much expert judgment is needed to keep the forecasts on track, and to do policy analysis. This judgment must be exercised through the back door, when intercepts get changed, or it is decided to use one equation from one model and another from another model. The quality of this kind of judgment is very hard to appraise.

At a more basic level, big models need lots of data, often on variables which are not measured. As a result, the models are often fitted to synthetic data of doubtful quality. Furthermore, the models involve the introduction of many somewhat arbitrary technical assumptions, about the functional forms of the equations, and about the behavior of the stochastic disturbance terms. To get the model going, it is usually necessary to compromise on the statistical niceties.

The subjective adjustments, the synthetic data, the technical assumptions, and the statistical compromises proceed to interact in bewildering ways, and the documentation just adds another layer of confusion.

3. An example

It may be useful to illustrate some of these points on RDFOR (Reduced Form Demand Forecasting Model). This model played a key role in the 1977 and 1978 midterm demand forecasts.

3.1 An overview of RDFOR[1]

RDFOR computes a demand surface for use in MEFS (Midrange Energy Forecasting System); this surface foretells what demand would be in e.g. 1990 for each type of fuel by consumption sector and geographical region, as a function of sector- and region-specific 1990 prices. The demand surface has a constant matrix of own- and cross-price elasticities. Thus, the demand surface can be defined by specifying one point on it, together with a matrix of elasticities.

RDFOR has three component parts:

(a) There is a system of log-linear demand equations, whose parameters have been estimated using econometric regression techniques from historical data.

(b) There is a procedure which employs these demand equations to fix one point on the forecasted 1990 demand surface.

(c) There is another procedure which employs the same demand equations to derive the matrix of own- and cross-price elasticities defining the shape of the forecasted 1990 demand surface.

3.2 The demand model

The demand model--point (a) above--is the structural heart of RDFOR. It consists of equations which set the log of quantity of fuel "demanded" (i.e., apparent consumption) equal to a linear function of the log of price and other variables. The RDFOR demand model recognizes the ten DoE regions, the four consumption sectors of the Annual Report (residential, commercial, industrial, transportation), and 13 types of fuel (coal, natural gas, electricity, gasoline, distillate, residual oil, etc.). Estimates of the parameters are made for each sector and region. The total fuel demanded in each sector and region is predicted first, as a function of the average price of energy for that sector and region and certain other explanatory variables (like population and income). The total is then shared out to individual fuels, with the share for each fuel depending on the price of that fuel relative to the average price of energy.[2] Fuel prices are exogenous to RDFOR, although they are endogenous to MEFS as a whole.

---

[1] This section is based on the reports by Parhizgari, and on interviews with EIA personnel. However, with respect to some critical issues, I have had to guess how things were done, due to the ambiguities in the documentation.

[2] Strictly speaking this discussion applies only to the "major" fuels in the residential, commercial, and industrial sectors. "Total" is something of a misnomer: RDFOR works with divisia indices. The transportation sector is different, but its structure cannot be deduced from the documentation. There is also a "fifth" factor, for feedstocks and the like. These are treated as "minor" fuels, and demand for them does not seem to be considered price-elastic.

The system of equations in RDFOR is large because of the extensive regional, sectoral, and product detail. There are many minor differences in the specifications of the various equations in the model. However, the equations used in the residential, commercial, and industrial sectors are all similar and identical in form across the ten DoE regions. I focus on the equations for total demand and pass over the equations used to share total demand out among the individual fuels.[1] The equation for total demand by a sector (residential, commercial, industrial) in DoE region $r = 1,\ldots,10$ and year $t = 1961,\ldots,1977$ is

$$(1) \quad q_{rt} = a_r + b_r p_{rt} + c_r y_{rt} + d_r h_{rt} + e_r c_{rt} + f_r q_{rt-1} + c'_r y_{rt-1} + d'_r h_{rt-1} + e'_r c_{rt-1} + v_{rt}$$

where

q_{rt} is the logarithm of an index of fuel consumption,

p_{rt} is the logarithm of a fuel price index,

y_{rt} is the logarithm of permanent income per capita,

h_{rt} is the logarithm of heating degree days,

c_{rt} is the logarithm of cooling degree days,

v_{rt} is a stochastic disturbance, and

a_r, b_r, c_r, d_r, e_r, f_r, c'_r, d'_r, e'_r are parameters to be estimated.

More specifically, in the residential and commercial sectors,

$$(2) \quad q_{rt} = \log(Q_{rt}/N_{rt}) ,$$

where N_{rt} is population, and Q_{rt} is a divisia index of fuel consumption. Similarly, in the industrial sector,

[1] For the 1978 report, the sharing-procedure was as follows. In the residential sector, the Hirst-Carney model was used, but its elasticities were constrained to match regional elasticities computed from RDFOR. In the commercial sector, RDFOR was used to estimate total demand. Then the national-level Jackson model was used to share this demand out among electricity, natural gas, and "oil". Finally, RDFOR was used to share the demand for these fuels down to the regional level, and to split "oil" into distillate and residual. In the industrial sector, RDFOR was used to estimate total demand, as well as the demand for electricity and liquid gas. The (undocumented) IFCAM model was used to share the remaining demand out among natural gas, distillate, residual, and coal. The main difference I can see between this and the proverbial back-of-the-envelope is that the old-fashioned method is self-documenting. After all, when you've finished doing the calculation, you've still got the envelope.

(3) $q_{rt} = \log(Q_{rt}/V_{rt})$,

where V_{rt} is value-added in manufacturing; the income coefficients c and c' in equation (1) are constrained to zero. For all three sectors,

(4) $p_{rt} = \log(P_{rt})$,

where P_{rt} is a divisia price index. "Permanent" income is a three-period moving average, with weights 4/7, 2/7, and 1/7. Money variables are in constant dollars.[1]

The stochastic disturbance terms in (1) are taken to be autoregressive, the parameter λ_r depending on the region:

(5) $v_{rt} = \lambda_r v_{rt-1} + w_{rt}$

Here, the 10-vectors $(w_{rt}: 1 \leqq r \leqq 10)$ are assumed to be stochastically independent of one another, prices, incomes, weather variables, and quantities consumed in previous years. The distribution of $(w_{rt}: 1 \leqq r \leqq 10)$ is assumed to be constant over time, with mean 0, but the 10×10 covariance matrix is not constrained. These assumptions are not defended in the documentation. In fact, they are not even spelled out.

This completes the specification of (1), with one major lacuna. The coefficients in (1)--but not the intercept a_r, or the autoregressive parameter λ_r in (5)--are constrained in RDFOR to be equal, or to vanish, across arbitrary groups of regions. In the residential sector, for example, the coefficients in (1) are constrained to be equal on the following "super-regions":

$$\{1,2\}, \{3\}, \{4\}, \{5,6,7,8,9,10\}$$

This was for total residential energy use. By way of comparison, in the corresponding equation for residential use of electricity, the super-regions are

$$\{1,2\}, \{3\}, \{4\}, \{5\}, \{6,7\}, \{8\}, \{9,10\}$$

So RDFOR considers that regions 5 through 10 behave the same when it comes to total residential energy use, but differently when it comes to turning on the electricity!

---

[1] This crucial fact is not stated in the documentation--the SYNERGY reports by Parhizgari. Likewise, from the documentation, it is impossible to tell whether the three-period moving average is applied to income, income per capita, log income, or log income per capita. I believe it is applied to income per capita. The functional form of the industrial demand equation is not specified in the documentation.

At the risk of anticlimax, in the equation for total commercial energy use, the coefficients are constrained to be equal across all ten regions. But when it comes to commercial use of electricity, different super-regions are used for different variables. For instances, regions 9 and 10 are required to show the same response to price changes, but are also required to show different responses to income changes: the coefficient for income is constrained to vanish in region 9, but not in region 10.[1] The RDFOR documentation does not even make these constraints explicit, let alone justify them.

Once the specification is completed, i.e., the linear constraints on the coefficients have been imposed, the system is estimated with the regression package TSP (Time Series Processor). This uses the Cochran-Orcutt method to handle the autoregression, and Zellner's method to handle the inter-regional covariances. These are both iterative procedures of the "generalized least squares" type. In effect, each equation in (1) is fitted separately by ordinary least squares. Then the residuals are examined, to estimate the autoregressive parameters and the inter-regional covariances in (5). Then the data are transformed to remove the estimated autoregressions and covariances and to equalize the variances. The transformed data are used to re-estimate the equation and the revised results are used to re-estimate the autoregression and covariances. The process is repeated until some standard of convergence has been met.

The equation system was estimated using data contained in the FEDS (Federal Energy Data System) data base.[2] This source contains most of the annual data required by the model for the period 1960-1977. The fitting period, however, runs from 1962-1977: two years of data are lost due to the lags.

3.3 Forecasting with RDFOR

The forecasted 1990 demand surface used in MEFS has constant elasticity, so it can be defined by specifying one point on it, together with a matrix of elasticities. The specified point is of the form (Q,P), where Q is a vector of quantities, and P a vector of prices. The P is simply the vector of 1990 prices determined by the last MEFS run on a comparable scenario. The Q is the forecasted vector of fuel quantities to be demanded in 1990 at prices P, computed from the RDFOR demand equations like (1) above.

Of course, to implement this idea it is necessary to forecast the exogenous variables in the demand equations, for every year between the present and 1990. For prices, this is done by a linear interpolation between the

[1] This discussion applies to the 1977 version; in 1978, quite a different pattern of constraints was imposed!

[2] The RDFOR documentation does not define its variables in terms of FEDS variables. In fact, it does not even specify FEDS as the data source.

current (e.g., 1977) prices and the 1990 prices P described above.[1] This interpolation defines the "price path". Some of the other exogenous variables in the demand equations (1), like population and income, are projected into the future by other agencies: the Census Bureau for population,[2] Data Resources Incorporated for income.[3] The variables describing the weather[4] are dropped from the demand equation (1). The intercepts a_r in (1) are now quite a bit off, but they are discarded and replaced by new estimates derived from "benchmarking": in effect, the intercepts of the demand equations are adjusted so that, for the latest year in which data are available, energy consumption in that year estimated from the equations matches the data exactly.[5]

The modified demand equations can then be iterated out to 1990, along the price-path and the trajectory of forecasts for the other exogenous variables: iteration is needed due to the presence of the lagged variable q_{rt-1} on the right hand side of (1). The resulting projection Q of quantities demanded in 1990, and the assumed 1990 prices P, define the (Q,P)-point on the 1990 demand surface, and hence the position of that surface. In some runs, the position of the 1990 demand surface is changed by EIA forecasters. These adjustments are called "demand offsets" and are used to model the replacement of traditional fuels by solar and geothermal sources of energy and, as the documentation says, the "exogenous reduction in the underlying energy demand that results from...conservation initiatives." This completes an overview of the procedure used to fix one point on the forecasted demand surface for 1990.

The next topic is the procedure used to define the shape of the demand surface: namely, its matrix of elasticities. The modified demand equations described above do have a constant matrix of own- and cross-price elasticities: constant over time, as well as over prices. However, this "short-run" matrix is not the one used in MEFS. Instead, RDFOR computes a matrix of "long-run" elasticities. These are constant over prices, but interestingly enough, they are time-dependent. Their computation will now be explained.

---

[1] This is the description in the documentation. However, EIA personnel inform me that the interpolation is log linear.

[2] Consumption is modelled <u>per capita</u>, as in equation (2).

[3] The Data Resources forecast is national. It is shared down to regions using 1974 shares derived from Bureau of Economic Analysis data. Thus, MEFS cannot predict changes in regional income shares, or relate such changes to energy variables.

[4] The count of gas service customers is dropped from the sharing equations, and the autoregression in (5) is also dropped. Dropping variables, of course, may introduce a substantial bias into the forecasts.

[5] The consumption data for the current year is "adjusted" to be what it would have been if the weather had been average. Even without this wrinkle, estimating the intercept from one data point is bound to cause large variances in the forecasts.

As described above, RDFOR makes a linear interpolation between current (e.g., 1977) prices for energy and the 1990 prices P derived from a previous MEFS run, to define the price path; then RDFOR iterates out along this price path using the modified demand equations to predict the quantities Q of energy demanded in 1990, by sector and region. To get the elasticities, the process is repeated, with a ten-percent increment in the 1990 prices. The difference quotient for 1990 quantity against 1990 price is used to approximate the long-run elasticities.[1]

3.4 A critique

Subjective adjustments are rampant. For example, the intercepts a_r of the demand equations (1) are changed to reflect the supply of exotic fuels, or the impact of conservation measures. In 1978, RDFOR did not predict reasonable fuel shares, so sharing was accomplished by splicing three unrelated models into RDFOR: namely, the Hirst-Carney model of the residential sector, the Jackson model of the commercial sector, and the IFCAM model of the industrial sector.

What about data? Most of the sectoral consumption data needed for RDFOR's equations simply does not exist. As a result, RDFOR was fitted to synthetic data, from FEDS. I will tell only three stories about FEDS. First, consumption data on distillate oil is collected by the Bureau of Mines. One consumption category is "heating oil". FEDS residential and commercial consumption data on distillate are constructed by sharing out this BoM heating oil. The shares vary by state, but not by year. Thus, time trends in

[1] Here is the sort of objection that can be raised to this procedure. MEFS is based on an "Economics 1" picture of market clearing. Thus, the quantity of fuel demanded in e.g., 1990 should be a function, in 1990, of 1990 prices. And it is the 1990 prices which will adjust to clear the market by moderating demand. Changes in 1987 prices (or 1993 prices) cannot help much with this. Thus, MEFS should be using the 1990 (short-run) demand surface from RDFOR to figure the 1990 equilibrium prices. However, MEFS is using a hybrid demand surface, whose position is determined by RDFOR's short-run demand equations (subject to adjustments described above), but whose shape is determined by RDFOR's "long-run" elasticities. Apparently, RDFOR has not been properly integrated into MEFS. For the 1978 Annual Report, the "equilibrium" prices from the integrating model in MEFS are fed back to the demand model, generating a new demand surface, which is then run through the integrating model, and so on. This iterative procedure continues until the system converges. However, the "long-run" elasticities are in effect built into the procedure used to forecast prices, namely the linear interpolation used to define the price path. In MEFS, market-clearing in e.g., 1990 influences prices in the years prior to 1990. This is a disturbing logical point: the direction of causality is reversed. And from an empirical point of view, recent experience suggests that energy prices change in drastically non-linear ways. These issues are not taken up in the RDFOR documentation.

sectoral shares--surely of considerable policy interest--do not appear in the data. Separate price elasticities are estimated in the residential and commercial demand equations for distillate. Clearly, however, the data does not support such an endeavor.

Second: FEDS consumption data on electricity are derived from the Edison Electric Institute. But Edison Electric reports "commercial and industrial" use together, in the aggregate only. Within this category, consumption is broken out separately for "large" and "small" users, although the break-point is undefined. FEDS takes the "large" users for its industrial consumers of electricity, and the "small" ones as its commercial consumers. Apparently, when Macy's turns on the lights, it goes into FEDS as an industrial concern.

The third example is on price data: specifically, the price of liquified gas. The data source is Platt's Oilmanac, which collects price quotes in only 11 cities. FEDS gives a price for each of the 50 States. How is this done? The idea is to surround each State, to the extent possible, by some subset of the 11 cities with price quotes. Then the average over the subset is imputed to the state.

I turn now from data to the logic of the model itself--the equations. These equations are not derived from economic theory, or from a detailed knowledge of the fine structure of the energy market. They represent a set of simplifying assumptions, needed to get on with the job of estimating and forecasting. So we have to ask whether these assumptions are reasonable, and what their implications are for the forecasts.

One key hypothesis in equation (1) is that price elasticities are constant. This may be a reasonable first-cut description of the energy market during RDFOR's fitting period (1962-77). But as the economy moves from a period of energy abundance to a period of energy scarcity, this hypothesis makes less and less sense. As the cost-share of energy goes up, its price elasticity should change. A forecasting procedure which assumes constant elasticity is likely to be too pessimistic about the possibilities of substituting capital, labor, and technology for energy. The technical assumption of constant elasticity is exerting an influence on the forecasts, and this influence gets stronger as the scenarios diverge from the circumstances which obtained during the fitting period.

Another assumption embedded in equation (1): in the industrial sector, the dependent variable is quantity demanded per dollar of value added, which responds only to prices and the weather. In particular, with prices and the weather held constant, the demand for energy is independent of the industrial mix. Again, this may be a reasonable first-cut description of the energy market during the fitting period. But as a statement about the future it is unrealistic, since energy is so widely used as a factor of production. A substantial change in the price of energy will sharply disturb the entire structure of relative factor prices. Some products take relatively little energy to manufacture, while others require relatively large amounts. If the relative price of energy goes up, U.S. industry should be substituting energy-efficient products for energy-inefficient ones: in other countries

this already seems to have happened. The equations do not contemplate such substitutions, and this make make RDFOR too pessimistic about the impact of rising energy prices on the economy, just as the assumption of constant elasticity did.[1]

I will bypass some aspects of the specification, like the arbitrariness of the inter-regional constraints,[2] or the crudeness of the assumptions about the stochastic disturbance terms.[3] However, some of the fine structure does call out for attention. For example, why does equation (1) use a divisia index for quantity and price, rather than total btu's and btu-weighted average price? Why have lagged weather variables, and three-period moving averages of income? In the nature of things, these questions are unanswerable; the hope is that such choices do not matter. However, in RDFOR they seem to matter a lot. In preliminary experiments with RDFOR-like equations, dropping the lagged income and weather variables tripled the estimated long-run own-price elasticity. These equations were fitted using btu-totals for quantities, and btu-weighted averages on prices. Moving to divisia indices tripled the elasticity again.[4] Technical assumptions matter.

---

[1] Over the fitting period, Q/V and P show little inter-temporal variation, except for the post-embargo run-up in prices. There is cross-sectional variation, but this is confounded with the effect of omitted variables. After all, the ten DoE regions differ among themselves in important ways other than energy prices and the weather. The mix of industries differs from region to region, influencing efficiency--Q/V on the left hand side of equation (1). The industrial mix also affects the distribution of the demand for energy among different fuels--and hence the price index P on the right hand side, for this represents an average over the different fuel types. The coefficients in RDFOR's industrial demand equation are heavily influenced by the cross-sectional variation in the mix of industries. This equation is therefore not a reliable guide to the behavior of these regions over time, in response to exogenous price changes.

[2] For example, in the industrial total demand equation, the coefficients are constrained to be equal across all ten DoE regions. Since the industrial mix is so different from region to region, this constraint violates economic common sense.

[3] Correlations are allowed across regions, but not across fuels or sectors. This is rationalized on the basis of "measurement error" in the data. But the imputations used to create FEDS are unlikely to correlate errors across regions; they are very likely to correlate errors across sectors. If the errors are correlated across sectors or fuels, the coefficient estimates in RDFOR may be compromised. Likewise, any remaining autocorrelation in the w_{rt} of equation (5) will become entangled with the lagged dependent variable in equation (1), and bias may result in the estimated speed of adjustment.

[4] For some discussion of the impact of choice of indices, see the papers by Hausman or Nguyen-Barnes.

Why is RDFOR so sensitive to minor changes in the specification? One reason may be that the equations are over-fitted. By way of example, take equations (1) and (2) for total energy demand in the residential sector.[1] The coefficients in this equation are constrained to be equal within four "super-regions", as discussed in section 3.2 above. However the intercepts a_r in (1), and the autoregressive parameters λ_r in (5), are left free to vary across all ten DoE regions. Table 1 below shows the count of parameters.

Table 1. Parameters in the RDFOR residential total demand equation.

coefficients	4×8	= 32
intercepts		10
autoregressive parameters		10
parameters		52

The data runs from 1960-77, but 1960-61 are lost due to lags, leaving 16 years for estimation. There are 10 regions, and $16 \times 10 = 160$ data points, so there is hardly enough data for estimation. The excess of parameters is aggravated by the fact that there are 10 variances and 45 covariances for the regional stochastic disturbances--the w_{rt} in equation (5). These have to be estimated too.[2] The residential total demand equation is very short on data, as shown by table 2 below.

Table 2. RDFOR is over-fitted. A count of things to be estimated, compared to the data points.

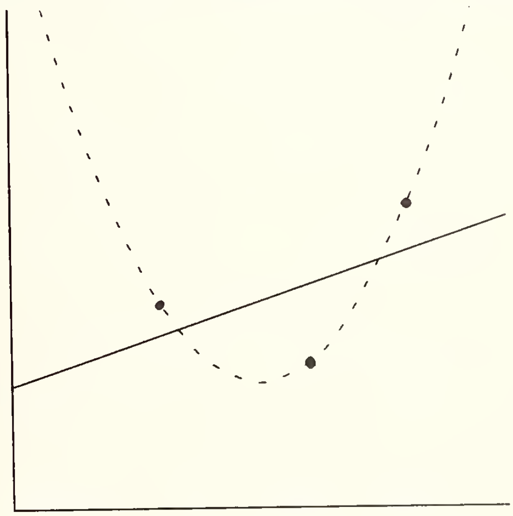
parameters	52	
variances and covariances	55	
things to estimate	107	
versus data points		160

[1]The residential sector is the worst.

[2]Asymptotically, the variance-covariance matrix can be well estimated, so degrees of freedom are not lost by introducing it. However, RDFOR's sample is quite finite. The 55 variances and covariances cannot be accurately estimated from 160 data points. So the rotated data are likely to involve dependencies. As a result, the coefficient estimates are likely to show large random errors. Their standard errors are computed on the basis of independence (as is asymptotically correct), and these may be misleadingly small. Finally, since the rotation matrix is random, the estimators may be quite biased. Generalized least squares is unsatisfactory, when there are so few data points relative to the number of parameters.

RDFOR's many parameters allow it to track the historical data with artificial and misleading precision. Small errors in the data, or small revisions to the data, are likely as a result to have a disproportionate impact on the coefficient estimates. That is, the coefficient estimates are likely to be subject to large random errors.[1] And unless the specification is absolutely right, which seems implausible, specification error can introduce large biases. This is illustrated in figure 2 below. If the three points really follow the straight line, putting in an extra coefficient and fitting a parabola has very bad consequences. I think RDFOR is making exactly the same blunder, albeit in several hundred dimensions.

Figure 2. Over-fitting can cause serious bias.



The statistical instability in RDFOR's structure can be seen by comparing the coefficients in the 1977 and 1978 versions of the model. In the industrial totals equation, for example, the long-run price elasticity estimated in 1978 was triple the one estimated in 1977, due to the addition of one data point and the revision of another.[2]

3.5 Summary

In this section, I have reviewed one econometric demand model used in midterm forecasting for the 1977 and 1978 Annual Report, namely, RDFOR. The object was to illustrate the following points:

---

[1] The documentation does not report R^2's, Durbin-Watsons, or standard errors.

[2] The long-run elasticities were computed as the price coefficient divided by one minus the lag coefficient from the numbers reported by Parhizgari.

- The documentation is quite inadequate.
- Many subjective elements go into the forecasts.
- The data are largely synthetic and the imputations questionable.
- The model embodies many somewhat arbitrary technical assumptions, which may have a lot of leverage on the forecasts.
- The fitting procedure is not well justified.

4. Some empirical evidence

So far, I have not addressed the question of whether the models might work, despite their shortcomings. For energy models, I have no empirical evidence to offer, one way or the other. However, I can speak about demographic projections and economic forecasting. Something can be learned from experience in these other fields.

To begin with demographics, recent work by Keyfitz and Stoto indicates that:

- Complicated population models do no better at forecasting then simple ones.
- No models forecast well.

For example, take the Census "median" projections for the U.S. population over the period 1945-1975. The root-mean-square error in the forecasted growth rates is about 5 per 1,000 per year.[1] Over the same period of time, the observed annual growth rates ranged from a low of 7 per 1,000 to a high of 19 per 1,000. As a result, a confidence interval for the forecasts of plus-or-minus one r.m.s. error is comparable in length to the entire range of historical experience for the period.[2] Clearly, population forecasting is a hazardous experience. Keyfitz gives only the following advice: quote a 68%-confidence interval around the forecast, a 95%-interval will be too discouraging.

The record in economic forecasting is difficult to read. There are many models, each forecasting numerous variables, over different time horizons. There are various measures of accuracy, and several conventions for handling data revisions. However, McNees (1980) offers a useful comparison of the accuracy of forecasts made by the groups at Chase, Data Resources Incorporated, General Electric,[3] and Wharton over the period 1970-79.

---

[1] The look-ahead period ranges from 5 to 30 years. In e.g., 1945 we have forecasts for 1950, 1955, 1960, 1965, 1970, and 1975. In 1970, only the forecast for 1975 is used.

[2] The record for other forecasting agencies and other countries is very similar. The results do not seem to depend much on the look-ahead period.

[3] General Electric does not use a simultaneous-equation model. This group uses some identities, some econometric equations, and a lot of intuition. Such mixed forecasting systems will be discussed again below.

The track records are all very similar.

A useful measure of accuracy is the Theil coefficient, which compares the root-mean-square error in forecasting changes with the root-mean-square of the changes themselves. For quarterly changes over the 1970's, the Theil coefficients in table 3 below are representative. Thus, a typical forecast for the quarterly change in real GNP will be in error by about 50% of the typical quarterly change in that quantity. Errors in forecasting real GNP and in forecasting the price level tend to go in opposite directions, so the errors in forecasting nominal GNP are smaller. All in all, econometric models do better than persistence forecasting,[1] but not by as much as might be expected--especially on volatile series, like housing starts or bill rates.

Table 3. Some Theil coefficients in forecasting quarterly changes[2] in the 1970's.

Real GNP	50%
Price level	30%
Nominal GNP	20%
Housing starts	90%
Treasury bill rates	90%

Zarnowitz (1979) presents comparable results for a somewhat earlier time period. He also makes an interesting comparison between judgmental and econometric forecasters. There turns out to be little difference in track records. Many other authors have come to similar conclusions. However, such studies have a common flaw: individual econometric forecasters are usually compared with the median judgmental forecaster. In principle, this could tilt the balance in favor of the judgmental forecasters, by letting their chance errors cancel out. Zarnowitz recomputed the comparisons, first averaging the absolute errors over time for each of his judgmental forecasters, and only then averaging across forecasters.[3] The conclusion holds up: the forecasters who use econometric models do not seem to have any real advantage over the judgmental forecasters.

When using an econometric model, of course, it is necessary to forecast the exogenous variables. Most forecasters also find it necessary to make subjective adjustments to the intercepts. However, after the fact, an assessor could run the model using the observed values for the exogenous variables, and without making any subjective adjustments. Christ (1975)

---

[1] A persistence forecast is for no change.

[2] One to four quarters cumulative changes.

[3] Private communication. Apparently, due to the strong correlation among forecasters, using the median does not in fact distort the comparison.

reports that this can easily triple the average size of the errors. Now a model is useful in policy analysis only if we believe it captures some economic structure, so the equations describe stable relationships among the variables of interest. To assess such a claim, it is the forecasts made with the observed values of the exogenous variables, and without the subjective adjustments, which seem relevant. Christ's results suggest that someone can make quite good forecasts based on econometric model, without establishing much of a case for its usefulness in policy analysis.

Even if the absolute predictions are quite inaccurate, it is often argued that a model may be useful in policy analysis, for comparing scenarios. What matters is not the absolute predictions, but relative comparisons. Christ reports the results of policy simulations, using several models whose absolute predictions agreed closely among themselves and were in reasonable agreement with the data. Impact multipliers--relative comparisons--for fiscal and monetary policy were computed from the different models and showed little agreement.[1] As Christ says,

> ...Though the models forecast well over horizons of four to six quarters, they disagree so strongly about the effects of important monetary and fiscal policies that they cannot be considered reliable guides to such policy effects, until it can be determined which of them are wrong in this respect, and which (if any) are right.

Finally, I report two provocative conclusions from Ascher (1978):

- Econometric forecasts have not been improving over time.
- There is no systematic advantage to big models, or to little ones, in forecasting accuracy.

5. Recommendations

- More effort should be invested in the collection of data for analytical purposes: part of the task here is to decide what kind of data are needed.
- More effort should be invested in quality control and in documentation.
- Forecasts should be made in much less detail.
- Simpler, more robust, and more stable forecasting technologies should be developed.

I make these suggestions in the belief that many questions do not have answers, and that all concerned would do well to recognize this. EIA may even have a responsibility to educate its clientele on this score. Furthermore, I think that there is of necessity a large subjective component in any forecasting exercise. Frank recognition of these facts may enable EIA to

[1] In part, this is due to the arbitrariness of the identifying restrictions; see Lucas and Sargent (1979).

dispense with a lot of complexity, allowing judgment to be brought to bear openly and directly on the important issues. This could enhance the credibility of the forecasts--and might improve their accuracy.

I will now sketch three alternatives to forecasting with big models, each based on a different view of the statistical realities, and then make a general comment.

Statistical engineering. On this view, there are no stable structural relationships among the economic quantities of interest. However, there may be fairly stable empirical correlations which could be discovered and used for forecasting. The regression equation is just a smooth estimate for some conditional expectation and is not a structural relationship: thus, policy analysis by regression equations is not feasible.

Constrained econometrics. The premise is that a stable economic structure underlies the data, but there aren't enough data to estimate this structure. There are too many goods, prices, and regions, and not enough years. The solution is to constrain some of the key elasticities to *a priori* values or intervals and make the estimates subject to these constraints.

Mixed forecasting. The idea here is to use judgment, guided by observed empirical regularities and *a priori* knowledge, to forecast the main aggregates. If necessary, such forecasts could then be shared out (e.g., to regions) by some mechanical regression procedure. An eminent Government statistician of my acquaintance has recently proposed the concept of "ignorance-oriented modelling". The terminology is pretty awful, but I think the idea is a good one. Basically, it is to identify the major uncertain factors affecting the forecast which is made in some simple and explicit way depending on these factors. The uncertainty can then be traced through to the final result. His kind of projection is one example of the sort of technique I have in mind.

Here is another example. When I worked at the Bank of Canada in 1970, they had a very good judgmental forecaster. used the traditional smoke-filled-room method to forecast the main aggregates. But then, he would run the numbers through a computer program. This program did the routine arithmetic, verified the accounting identities, and checked on the plausibility of some key relationships among the aggregates: for instance, the savings ratio had to fall in some reasonable range and so did labor productivity. If the numbers didn't balance, they went back to the smoke-filled room for repairs. This system seemed to me to have many of the desirable features claimed for econometric models: the numbers had to add up and be consistent with the past.[1]

[1] Still another mixed forecasting system is the one used by General Electric. As discussed above, these forecasts are quite comparable in accuracy to the ones from large models at Chase, Data Resources and Wharton.

A general comment. The three approaches sketched above are quite different, but do have one thing in common: they would be far simpler to use than the present forecasting technology. This seems to me to bring real advantages. Using a simpler forecasting technology would free up resources for collection of data relevant to modelling, for quality control, and for documentation. It might even give EIA analysts some time to do analysis in-between the annual convulsions over the report to Congress.

Appendix A. Over-fitting regression equations

In order to demonstrate the pitfalls in over-fitting a regression equation (i.e., having too few data points per parameter), the following experiment was performed. A matrix was created with 100 rows (data points) and 51 columns (variables). All the entries in this matrix were independent observations drawn from the standard normal distribution.[1] In short, this matrix was pure noise. The 51st column was taken as the dependent variable Y in a regression equation; the first 50 columns were taken as the independent variables $X_1,\ldots,X_{50}$. By construction, then, Y was independent of the X's. Ideally R^2 should be insignificant by the standard F-test. Likewise, the regression coefficients should be insignificant by the standard t-test.

These data were analyzed in two successive multiple regressions.[2] In the first pass, Y was run on all 50 of the X's, with the following results:

- $R^2 = 0.58$, $P = 0.036$;
- 21 coefficients out of 50 were significant at the 25% level;
- 5 coefficients out of 50 were significant at the 5% level.

Only the 21 variables whose coefficients were significant at the 25% level were allowed to enter the equation on the second pass. The results were as follows:

- $R^2 = 0.51$, $P = 5 \times 10^{-6}$;
- 20 coefficients out of 21 were significant at the 25% level;
- 13 coefficients out of 21 were significant at the 5% level.

The results from the second pass are misleading indeed for they appear to demonstrate a definite relationship between Y and the X's, that is, between noise and noise.

This simulation captures two features of much empirical work, including the statistical work that went into RDFOR:

- The ratio of data points to parameters is low.
- Variables with small coefficients are dropped and the equation re-fitted without them.

This simulation shows that such practices can produce highly misleading results.[3]

---

[1] Simulated on the computer.

[2] Using SPSS. When the data were passed to SPSS, five variables were inadvertently set equal to zero and did not enter the regression for that reason. SPSS allows intercepts in regression, not counted here as coefficients; in both instances, the intercept was not significantly different from zero.

[3] Asymptotic calculations support the simulation results, and will be reported separately. Also see section 2 of Huber (1973), as well as Rencher and Pun (1980).

References

J. S. Armstrong (1978a). Forecasting with econometric models: folklore versus facts. Journal of Business, Vol. 51, pp. 549-600.

J. S. Armstrong (1978b). Long-range Forecasting: From Crystal Ball to Computer. Wiley, New York.

W. Ascher (1978). Forecasting: An Appraisal for Policy-makers and Planners. Johns Hopkins, Baltimore.

C. Christ (1975). Judging the performance of econometric models of the U.S. economy. International Economic Review, Vol. 16, pp. 54-74.

R. Fair (1979a). An analysis of the accuracy of four macroeconometric models. Cowles Foundation Discussion Paper, to appear in the Journal of Political Economy.

R. Fair (1979b). Estimating the expected predictive accuracy of econometric models. Cowles Foundation Discussion Paper, to appear in the International Economic Review.

D. Freedman (1979). An assessment of the READ model. Validation and Assessment Issues of Energy Models, National Bureau of Standards special publication 569, ed. by Saul Gass.

D. Freedman (1980). On uncertainties in model forecasts. Technical report, Lawrence Berkeley Laboratory. In progress.

D. Freedman, T. Rothenberg and R. Sutch (1980a). An assessment of the Federal Energy Data System. Technical report, Lawrence Berkeley Laboratory.

D. Freedman, T. Rothenberg and R. Sutch (1980b). The demand for energy in the year 1990: An assessment of the Regional Demand Forecasting model. Technical report, Lawrence Berkeley Laboratory.

A. Goldberger, A. L. Nagar, and H. S. Odeh (1961). The covariance matrices of reduced-form coefficients and of forecasts for a structural econometric model. Econometrica, Vol. 29, pp. 556-573.

J. Hausman (1975). Project Independence Report: An appraisal of U.S. energy needs up to 1985. Bell Journal of Economics, Vol. 6, pp. 517-551.

P. Huber (1973). Robust regression: asymptotics, conjectures, and Monte Carlo, Annals of Statistics, Vol. 1, No. 5, pp. 799-821, especially section 2.

D. Kahnemann and A. Tversky (1974). Judgment under uncertainty: heuristics and bias. Science, Vol. 185, pp. 1124-1131.

E. Kuh and J. Neese (1979). Parameter sensitivity and model reliability. Technical report, Center for Computational Research in Economics and Management Science, M.I.T.

E. Kuh and R. E. Welsch (1979). Econometric models and their assessment for policy. Validation and Assessment Issues of Energy Models, National Bureau of Standards special publication 569, ed. by Saul Gass.

W. Leontieff (1971). Theoretical assumptions and nonobserved facts. American Economic Review, Vol. 61, pp. 1-7.

R. E. Lucas and T. J. Sargent (1979). After Keynesian Macroeconomics, Federal Reserve Bank of Minneapolis, Quarterly Review, Spring.

S. Makridakis and M. Hibon (1979). Accuracy of forecasting: an empirical investigation. JRSS Ser. A, Vol. 142, pp. 97-145.

S. McNees (1973). The predictive accuracy of econometric forecasts. New England Economic Review, Federal Reserve Bank of Boston.

S. McNees (1974). How accurate are economic forecasts? New England Economic Review, Federal Reserve Bank of Boston.

S. McNees (1975). An evaluation of economic forecasts. New England Economic Review, Federal Reserve Bank of Boston.

S. McNees (1977). An assessment of the Council of Economic Advisers' Forecasts of 1977. New England Economic Review, Federal Reserve Bank of Boston.

S. McNees (1979). The forecasting record for the 1970's. New England Economic Review, Federal Reserve Bank of Boston.

J. Mincer, ed. (1969). Economic Forecasts and Expectations. NBER, Columbia University Press, New York.

O. Morgenstern (1963). On the Accuracy of Economic Observations. 2nd ed. Princeton University Press, Princeton.

L. E. Moses (1980). One statistician's observations concerning energy modelling.

H. D. Nguyen and R. W. Barnes (1979). An evaluation of the Midrange Energy Forecasting System. Oak Ridge National Laboratory, EIA/DOE Contract No. W-7405-ENG-26.

A. Parhizgari (1978a). Final documentation report on the Regional Demand Forecasting Model 1977 and 1978 versions. SYNERGY, Washington, DOE Contract No. EC-77-C-01-8560.

A. Parhizgari (1978b). Final report: Regional Demand Forecasting and Simulation Model User's Manual. SYNERGY, Washington, DOE Contract No. EC-77-C-01-8560.

A. Parghizgari (1979). Draft documentation report on the Demand Analysis System. SYNERGY, Washington, DOE Contract No. EC-78-C-01-8560 MOD(20).

A. C. Rencher and F.C. Pun (1980). Inflation of R^2 in Best Subset Regression, Technometrics, Vol. 22, No. 1, pp. 49-53.

H. O. Stekler (1970). Economic Forecasting. Praeger, New York.

V. and J. Su (1975). An evaluation of ASA/NBER Business Outlook Survey Forecasts. Explorations in Economic Research, Vol. 2, pp. 588-618.

V. Zarnowitz (1979). An analysis of annual and multiperiod quarterly forecasts of aggregate income, output, and the price level. Journal of Business, Vol. 52, pp. 1-34.

MODEL CREDIBILITY: HOW MUCH DOES IT DEPEND UPON DATA QUALITY?

C. Roger Glassey
Harvey J. Greenberg
Energy Information Administration
Department of Energy
Washington, D.C.

Credibility, like beauty, is in the eye or perhaps the mind of the beholder. Credibility depends upon many things, and data quality is only one of them. Furthermore, the credibility of a model is something that does not exist in an absolute sense, but rather within the context of the particular use to which the model is being put. For example, the simple model that says $x(t)=x(0) - 1/2\ gt^2$ is quite credible for describing the trajectory of a cannonball dropped by Galileo from the Leaning Tower of Pisa. However, it is not a good model for the trajectory of a goose feather because it neglects the effects of air friction. We should focus our attention not on the question of credibility of models, but on the question of the credibility of analytical results. A model is merely that part of an analysis that has been captured in the form of a computer program, usually because it is repetitive and algorithmic.

The credibility of analysis results seem to be a prerequisite for their use in a decision-making process. It seems to us that an analysis is credible if one of three conditions hold: (1) we agree with the conclusions; (2) the analysis is accompanied by a simple, logical explanation that uses a credible behavioral theory; and (3) we understand what assumptions and data were used in the analysis and the process by which they were manipulated to derive the conclusions--furthermore, we believe that the assumptions are plausible and the data are of sufficiently high quality for the purpose. This last condition, you notice, is full of qualitative statements. For example, how well do we need to understand the methodology of the analysis? Do we need to understand it well enough so we can replicate it? How plausible is plausible enough for the assumptions? How accurate is accurate enough for the data? The less disposed we are a priori to accept the conclusions of the analysis, the more stringent will be our standards in answering each of these questions.

All model users hope that there is more correspondence between the model and the real world system of interest. However, in not all models is this correspondence equally strong. At the strongest end of the spectrum are the models that are mathematical representations of the natural law. For example, the system of differential equations that is used to calculate space vehicle trajectories has great predictive power and rests upon theory that has been developed over centuries of careful observation, experimentation, and testing. These models are sufficiently credible for decision-making purposes that vast sums of money and many human lives are entrusted to the accuracy of their results.

There are weaker models that claim to be only crude approximations to a small number of allegedly significant features of a complex system. Log-linear models for energy demand as a function of income and price, for example, fall into this category. Such models do not claim to represent natural law; they merely rest upon two apparently plausible assumptions: (1) if the price of a good increases, people will tend to buy less of it, other things being equal, and (2) if people have more money, other things being equal, they will tend to spend more. There is, however, as far as we know, neither empirical nor theoretical justification for the particular form of the equations used. It is as much a matter of convenience as anything else.

At the weakest end of the spectrum are models that are used primarily to maintain internal consistency in the process of scenario building. The simple accounting structure used in the Short-Term Integrated Forecasting System comes to mind. By forcing both historical data and projections, some of which are essentially subjective assumptions, into this accounting framework, we can at least guarantee that consumption equals net production plus imports minus stock change. We go to this trouble inspired by the belief that an internally consistent set of forecasts is more credible than a collection of inconsistent forecasts.

All of this discussion is to suggest that there are several sources of uncertainty in model output. For the weaker models, it may be that uncertainty about model structure, uncertainty about the effects of the omitted variables, may be the dominant source of output uncertainty, and the data uncertainties are relatively unimportant. When the models are based on scientific law, the accuracy of the output depends largely upon the accuracy with which the coefficients of the model can be estimated. That is to say, the data uncertainties then dominant the uncertainties in the model output.

We have all used, or at least heard, the expression "garbage-in-garbage-out (GIGO)." Generally, this means we cannot expect useful results if the data are "bad." This poses a dilemma: How should policies be evaluated? As scientitst, we all prefer the apparent rigor of using a model; as scientists we prefer to reject the use of models whose data requirements put us into GIGO. Here are two ways to deal with this dilemma:

1. Refuse to evaluate, stating the reason and specifying the data requirement; or,

2. Temper the model results with expert analysis.

In practice, option 1 is rarely fully exercised. If refusal is feasible, it is generally not accompanied by a full explanation of why the data are unacceptable. At least as often, option 1 is not feasible. In this paper, we consider: (1) how the dilemma may be avoided for some cases, and (2) tempering model results.

A major reason for the dilemma is the persistent separation of data collection and modeling. What data to collect and in what form are detached from modeling and analysis. There is virtually no opportunity for modelers/analysts to participate in data collection, even to the extent of stating present or anticipated data requirements. No notion of "cost-effectiveness" seems to have been applied to the data collection decisions. Instead, data are collected to measure compliance with a law, not to aid analysis.

Thus, a preventative measure for the future is to have something like a "Data Acquisition Administrator," analogous to a Database Administrator. The Data Acquisition Administrator is responsible to all users of data, including modelers/analysts. Presently, EIA does not have such a function to the extent where data requirements are specified with a measure of "value" to be traded-off against data acquisition and maintenance costs. There is, however, a move towards this form of cooperative assessment. The National Energy Information System is being developed with reviews from Applied Analysis. Also, the task force for assessing the Oil and Gas Reporting System includes representation of the Office of Applied Analysis. This now brings us to how model results may be tempered, and why this option is in concert with having a measure of value to influence data acquisition.

The quality of analysis pertains to how well we can answer a posed question, generally to evaluate some policy or anticipate a preventable disruption in our economy. Thus, we analysts affect the quality--not just the model we use as a tool. Credibility pertains to how well we report the quality, preferably by some measure of uncertainty. Our basic thesis is this: working with "soft data" intrinsically identifies its importance. One example of this is the prominent Oak Ridge Residential Consumption Model, built by Eric Hirst. During its development, Dr. Hirst had to estimate some of the data, which are now collected.

As we use an imperfect model, we learn which data are decisive, which are moderately important and which have little effect on certain questions. It is the ongoing use of such imperfection, accompanied by "exploratory modeling," that provides a measure of value to specific data requirements. Furthermore, if we do not use models, tempered by expert analysis, then we rely on unstated assumptions, inconsistent projections and an amalgamation of cause versus effect.

We have all experienced a Delphi method approach, relying on assembled experts to reach a consensus. It often does not work; it is almost always prohibitively time-consuming and it is virtually impossible to document. Moreover, such an exercise does not collect wisdom and pass it on to others for subsequent advancement. Models, evolving to remain relevant and improve the quality of analysis, inherently provide a record of collected wisdom, and it can be subjected to critical review--a form of validation/assessment--in order to make it better.

Thus, we propose that GIGO is an overused cliché, ignoring three basic facts: (1) we must respond, with or without a model; (2) we can institutionalize intellectual growth; and, (3) we can evaluate data requirements with a cost-effectiveness perspective. The bottom line, therefore, is that we should not reject a model's usefulness solely because the data are presently unreliable or nonexistent.

REMARKS ON FREEDMAN AND GLASSEY-GREENBERG PAPERS

Douglas R. Hale
Director
Validation Analysis
Energy Information Administration

It is late in the day; and I have to review the work of an Assistant Administrator. I hope to be brief and, on occasion, vague. Please don't press me on vagueness.

Roger and David Freedman have suggested that EIA energy models and forecasts have credibility problems because:

- o the data base often does not support the model's needs;
- o the models and their forecasts are poorly documented;
- o some models are not competent; and
- o the models are big and complex.

They also suggest that these problems can be partially attributed to deeper issues of EIA's structure and to the environment in which EIA operates.

Roger and David Nissen have stressed the fact that there is no well established process for the modelers to influence data collection. I agree this is a serious issue. I am willing to spend several years making sure that data collection supports analysis needs.

David has pointed out that a combination of excessive demands for forecasts and the size and complexity of the models have contributed to poor documentation and uneven quality. He goes so far as to suggest that less ambitious questions and simpler models "... might even give EIA analysts some time to do analysis ...." I've been a Government analyst for seven years and have served in some of the best analysis and policy shops in town. Now I'm a bureaucrat. If David has found a way to let Government analysts do analysis, I'll go straight.

David Freedman also notes that the skill mix is not appropriate to the task. He mumbled something about statisticians: I would have added economists. It may seem strange to raise the personnel issue as a serious organizational concern--it may seem more an issue for middle management. I disagree. Government hiring and firing rules, together with inflexible entering salary schedules, virtually ensure that the skill mix will be suboptimal. Moreover, having personally interviewed over 150 job candidates since October, I am convinced that the Government barely competes for superior, entry level talent in a large number of fields. Unless the Government changes, we will always face skill mix problems.

In general, I agree with their assessment of the particular credibility problems and their institutional basis. However, I would add that it is inherently more difficult to recognize a good model than to recognize a well designed bridge. I would also suggest that challenging EIA forecasts with real data is not trivial.

We have said there is a credibility problem. I think it is appropriate to describe how the work of the Office of Validation Analysis, in conjunction with the Office of Oversight and Access, approaches some of the specific problems we have discussed.

As you know, the Office of Applied Analysis has a quality control program to ensure the quality and accessibility of those models used to prepare published energy projections and analyses. This program focuses specifically upon energy analysis models in use and currently emphasizes improving model documentation.

The Office of Energy Information Validation is responsible for conducting independent evaluations of the actual limits of major energy models, and for making independent recommendations to the Administrator for improving individual models and the modeling process. The model evaluations typically build upon the documentation, verification and sensitivity analysis efforts conducted by OAA. I want to stress that our purpose in studying models is to improve EIA's forecasts: our prupose is not to embarrass bureaucrats. That is too easy a job. Our current assessments are directed toward answering seven questions. We think that clear answers to these questions are fundamental to public understanding and evaluation of these models.

1. Considered as a computer device, can the model be understood and used by third parties?

2. What are the model's fundamental mathematical properties?

3. What is the model's logical (physical, statistical, economic, engineering, etc.) structure? What is the proper domain of model applications?

4. What is the nature of the data needed to prepare forecasts? To test the model?

5. Are the individual specifications and assumptions supported by data and theory?

6. What can be said about the reliability or uncertainty of the forecasts?

7. Given the above, for what purposes is the model suited? How is it actually used?

We have three model evaluation projects in process: the Long-Term Economic Analysis Package, the Short-Term Integrated Forecasting System, and the most recent Sweeney Transportation Model.

We have found that these questions need to be addressed in the context of the model. For example, I'm having difficulty saying anything about the statistical validity of non-statistical, deterministic models like LEAP. Testing that model in its own terms is problematic because there are no explicit questions of statistical formulation and method: one cannot, for example, contrast estimation techniques. Testing against real data is confounded by the five-year gap between equilibrium solutions.

This emphasis on fundamentals - Does the model solve? Can solutions be replicated? Is it stable? Do the data make a difference? - enables us to examine issues that are both important and less subjective. However, these seemingly mundane issues of objective performance have proven extraordinarily difficult to resolve. Consequently, relatively little attention is currently being paid to model comparisons and other measures of overall model quality.

These efforts will not solve the credibility problem. I believe, however, that without this work we'll never raise the level of the debate above that which we heard today.

REMARKS ON FREEDMAN AND GLASSEY-GREENBERG PAPERS

David A. Pilati
National Center for Analysis of Energy Systems
Brookhaven National Laboratory
Upton, New York

I wholeheartedly agree with David Freedman's paper, but I suspect that David Nissen and I have a lot of disagreements at certain points.

Professor Freedman attempted to look at a general question but really provided a specific application by looking at the RD4 model and answering "Are Energy Models Credible?" He was only addressing one particular model, not the whole class of energy models.

He made a serious accusation as to the validity of that model, one that I don't think I can take lightly. However, judging from the comments of others, my perception may be a minority opinion. I was trained as an engineer and I think that engineers look at the world very differently than social scientists and particularly economists.

David Freedman also noted (and I don't believe this was in the formal paper) that there was another important perception that we should be looking at, too. This was the perception that institutional setting drives the technical details in the approach that is used. This hasn't been mentioned by any of the other discussants but it seems to me that, once recognized, it is a very important area. What is the solution?

Do you try to outguess the institutional framework and the institutional response to technological change in using new technologies, in particular, in using new modeling technologies? Or do you just live with the situation?

Although I strongly agreed with Professor Freedman's position, he did make two technical errors that deserve some discussion.

His paper berated large-scale economic models for their inability to predict discontinuities. For example, you can't predict embargoes. However, this problem exists for all modelers. There are no models of physical phenomenon that I know that can predict precisely when a discontinuity can occur. The inability to predict a discontinuity is a generic problem and should be understood as such.

At best, models can only tell you what the probability of a discontinuity occurring is. In the energy field, we have a lot of possible discontinuities that one can think of, all with (hopefully) low probability. However, no person or model can calculate the probabilities associated with these possibilities.

He accused RD4 of having weak theoretical underpinnings and then went on to say that one of these weaknesses was the assumption of constant price elasticity in the industrial demand area, an area that I am very familiar with. Dr. Freedman implied that as energy prices increase, price elasticities must obviously increase, too. That statement is based on technological faith and, given our existing technological base, this is not going to occur.

In fact, using the industry process models at Brookhaven, we have shown that the energy price elasticity declines as the energy price increases because you soon run out of known technological devices that can further reduce energy demand. There are two sides to this coin; one is the thermodynamic limitations to conserving energy, and the other is lack of foresight in R&D planning.

It is somewhat reassuring that the Long-Term Conservation Technology Office has been reopened. Someone mentioned it earlier this morning. Why in the world it is in the Storage Division I will never know.

The second paper by Glassey and Greenberg asked the question, "How Much Does Model credibility Depend on Data Quality?" And then they skirted the question and really talked about ways of improving data quality.

Models can be conceptually divided into structure and data. You cannot expect the numbers that come out to be any better than the numbers that went in, even if the structure was 100 percent correct.

However, it is certainly conceivable that models can have a lot of use even if the data are poor because they can display behavioral relationships of a particular system that increases one's intuition and understanding of the modeled system.

Let me cite an example from the Brookhaven industry process models which use some questionable data like any other models. The process models show that as you increase the investment tax credit on conservation technologies, you start saving energy and then things sort of level off, and all of a sudden (at a tax credit level of approximately 60 percent in the paper industry) you start saving a lot of energy.

The analysts ask, "Well, what in the hell is going on?" Detailed analyses of model results show that the tax credit causes premature retirement of equipment. Other people might call that "overdue" retirement for many of our production facilities. Sooner or later, many of today's inefficient plants will have to be shut down and replaced with modern energy-conserving production facilities. Rising energy prices and conservation tax credits can hasten the day for such changes.

The equation on the board (equation for a falling body with no friction, for people who are not trained in engineering) is not a model. A model has assumptions. An equation has assumptions. Galileo was no doubt aware of friction, but he did not know how to model it. Otherwise, why did he drop cannon balls instead of feathers?

And any good engineer would not write down this equation and say it is the equation of motion for a falling body. He would have several explicit assumptions. This is something that doesn't seem to exist in most policy models.

I think it is obvious to me from David Nissen's discussion and David Freedman and my point of view that there is a growing split between science and pragmatism. Those that are more of a scientific bent are going to become more and more frustrated because it is fairly obvious that there are going to be tighter monetary problems in doing this kind of work and that pragmatism is going to ultimately win out.

DISCUSSION ON FREEDMAN AND GLASSEY-GREENBERG PAPERS

DR. GREENBERG: Just a few thoughts that came to mind in listening to the discussants. It has to do with the way in which we misuse the computer in doing some of this work.

Let me begin with a discussion of when I went to FEA in 1976. Bill Hogan said it would be a lot of fun, which it was for a while.

I stayed on after the reorganization, through the new regime. There have been a lot of changes from 1976 to 1980. But, I'm looking at the two recent annual reports to Congress and the past National Energy Outlook report, along with other studies. There is not evidence that the quality of analysis that has come out in the past two years is better than the quality of analysis that came out the first two years I was at FEA. It does seem different, and it does seem much more subjected to peer review.

It would be very hard to go back to some of the things that took place then and try to impose on them this kind of review process or try to pick out those elements that could be called science. I think what Larry [Mayer] meant is that it is not binary. It doesn't have to be science or non-science. Parts of it could be subjected to scientific scrutiny, while other parts may not.

I think very little of it was subjectable in the earlier days, but if you measure outputs instead of inputs, it is not clear that we are doing better. It is clear we are doing things differently.

The best of the two eras can be combined in what I think is EIA's singlemost important mandate, which has not yet come to pass. That is, the development of a National Energy Information System. Here information is not to be taken synonymous with data.

I think, properly constructed, the quality of the analyses that took place in '76 and '77 could be regained if we didn't spend half our computing resources reformatting data. This is done to accommodate outside assessor requests for our system and the multitude of software systems to be used by them, and the proliferation of files that don't increase information but only obscure the core of information.

What was good about the earlier analysis--the speed of execution, the quick turnaround, the ability not just on the computer but the ability to capture the thought processes--enabled us to communicate with a hand-full of analysts and so on--and go quickly think through an issue. Also, the support that was leant to the enterprise that I think made the quality difference was better before than now, in my judgement. The interference that we get, that the computer is more of an obstacle than a help at the present, tends to be more discouraging than helpful.

Standards and guidelines and the talk about documentation tends to interfere rather than help. It used to be that documentation was viewed as an assist to communication. Until you get the computer doing the documentation for you, which we used to do to some extent, you can't have the documentation keep up with the modeling and the data. It is going to be out of phase by one or two years. It takes that long to do it by hand.

Ideally, you would like to be able to push a button and get all the tedious stuff done like documentation, and let the talented analysts spend their time thinking about the substantative matters, and let the computer work for them rather than the other way around. A lot of analysts spend more than half their time writing JCL's or other kinds of non-interesting tasks. One of the consequences of this is you don't retain very many high-quality analysts. They tend to go to work for some obscure bank or something like that, and don't tend to hang around to add to the quality of the analysis.

DR. GASS: Thank you. I would like to have a question from the floor. Please use a microphone and identify yourself.

MR. BARNES: I am Dick Barnes, from Oak Ridge.

There has been a lot of talk here today about acquiring data for analysis, with the intention of reducing the uncertainty of the analysis. It seems to me, however, that there are elements in our society or economy whose best interest lie in maintaining the uncertainty of these analyses. They don't want big brother up there looking at all their private data and drawing inferences and telling them what they should be doing in the future.

I am wondering if there has been any consideration of the problem of really trying to develop this information bank and to obtain all of the data necessary to prepare these forecasts and policies with the uncertainty that they think they need.

DR. GASS: Thank you. Doug Hale, do you want to try that one?

DR. HALE: We have certainly given the problem consideration. I am not aware that we have come to any solutions. One thing that I can say is that the financial reporting system which we are examining now is a system to provide the very, very detailed accounts from the energy companies.

That system may, in fact, be a test bed for how far we really can go. As a policy analyst, I caused it to happen that similar data were required or will soon be required of major smelters. I was told by an executive vice president of the largest copper consortium in the world that he will close down every one of his smelters before turning over such data.

I think the progress of FRS is going to go a long way to answering questions about the most highly sensitive data available from corporations. I am not convinced that our problems are all of that sensitive. I am having people running around trying to figure out what heating oil is selling for. That is a big problem.

Well David has told stories about the regional data. Most of that data is not of the trade secret variety. We have also been having some problems in coal data, the production data of coal.

Basically, I think there is a lot of mileage to be had from carefully thinking through what analysis needs in terms of data and then going after it. I really doubt that most of us are going to be as sensitive as we can.

CHAIRMAN GASS: Thank you. Another question from the floor? Yes, Fred?

MR. MURPHY: I am Fred Murphy from DOE. In the energy discussions today, in terms of worrying about synthetic data which are essentially data that are output (supposed data which are the output of the model), I have a basic question.

Is there such a thing as data, or are all data synthetic? It seems that a good example is EIA's completed first survey of the oil and gas reserves of the country. Now what is the information that EIA has there that is the output of a set of differential equation models that you ultimately use to transform to what they have under the ground?

I would like to address this to David Freedman. Is there anything other than synthetic data?

DR. FREEDMAN: Yes.

DR. GASS: Could you give us an example, David?

DR. FREEDMAN: I think, for example, when the Bureau of the Census goes out and takes the current population survey every month you are not getting 100 percent response. There is some imputation involved. But the data are a lot closer to the real end of the spectrum than when you take LPG prices in 11 cities and spread them out to different states.

That is, you know pretty much how the Bureau of the Census went out into the field and you know how they took their household sample. You also know the interview procedures they follow. You even know in some detail the imputation procedures they used and you even know their quality control studies on those imputation procedures.

MR. MURPHY: Now what you are arguing is a case of degree, not a case of difference. A true distinction.

DR. FREEDMAN: Sure.

DR. MAYER: I think what is important to say on that, Fred, is going back to my point about making important discoveries. What have we learned in last seven years about energy?

The important point is that in these other related areas--population, health care--it is not that their data were of good quality or bad quality, but is that whatever their data were they discovered something.

It seems to me that the real question is what are we discovering with all this money. Instead of the assessment side saying the modelers don't deserve theirs and the modelers saying the assessors don't deserve theirs, do any of us deserve ours?

DR. GREENBERG: I think we learned a little bit about the affects of price controls. I think we know a lot more now than we knew seven years ago about the impacts of oil and gas controls.

DR. MAYER: Well I would be interested--I will take what David Freedman coined if he doesn't mind the embarassment. That is, I would be interested in someone sending me some testable; empirical statements about the energy world that some of our modeling and analytic efforts are based on. Then I can go out and do something very simple, as I can in physics or chemistry and health care, I can test it.

Just like I had a chemstry set when I was a little kid, I could see that if you mixed X and Y, you blew up most of your bedroom. I would like to know what simple statements we have learned.

DR. NISSEN: I thought I had provided an example of that.

DR. MAYER: Did you provide a model or not, David?

DR. NISSEN: No. This stage of the discussion always puts my teeth on edge. For any of us who have actually sat in a room late at night with guys trying to figure out energy policy, the point is that the modeling effort is essential.

This is a silly kind of discussion. We said they are not going to build the damn generating plants because they don't need them. And we were right. They didn't need them.

DR. FREEDMAN: I would like to make a brief rejoinder to some of the comments that Dave Nissen made.

I guess the first thing is to make another public confession of failure. I guess I didn't convince him. There is one thing he said that I would really like to agree with, and that is let the data have a vote.

That seems to me to be a crucial idea and I fully endorse it. But there is, I think, a major difference between the two of us in that it is precisely my view that making large-scale mathematical models of the kind we have talked about today prevents the data from getting the vote.

What you want to do is you want to take the 10 numbers you have. You take the 10 numbers you have and look at them, inventing another 100 numbers and fitting a giant regression equation with 50 parameters by some complicated fitting procedure. I don't think that lets the data have a vote. That lets the model and the assumptions in the model vote.

Now I don't want to go on much longer. I want to ask a question of Dave and it is going to be different from the question Larry asked me because it is not rhetorical, it is a genuine one.

I want to preface it by reviewing your review of our debate. All right? I say the data are poor and you agree. I say the method of estimation is poor and you don't disagree. I mean you call it a quibble but you did not disagree with me.

I had some serious questions about the specifications of the model and you didn't mention that. And I also said that there is no track record of success in this game and you didn't really take that up either.

And now the question I want to ask you is why should we believe the results from the model?

DR. NISSEN: Why would somebody who wanted to make energy policy decisions use models which are different than in what they believe? You persist in evaluating the model as a piece of science. And I, again, persist in claiming that that shows an institutional ignorance of what models are used for.

Models are used to mechanize the image of the world that the decision maker has. You explain to him as simply as you can, "Look, this piece over here says that if you double the price, the demand is going to go down 60 percent. Do you believe that?"

He will say, "Well, should I?" Then you get into a discussion of whether or not it is believable. Or he says, "No. Run it with zero." We will say, "All right. That happens to be something that is probably a debatable scientific position but it is something that we will do for you in terms of assisting your evaluation of your choices."

So the problem isn't the believability of the model as a representation of reality. That is part of the problem but that is not the problem. The problem is does the model serve a use in organizing the available community of shared information or supposition about states of the world and modes of behavior; can the model then perform the arithmetic so that it calculates the mass balance stuff, that it gets the valuation equilibrium straight, and all that stuff which is hard to do in a 10-equation model. Or it is hard to do in your head. Does it come up with an internally consistent set of results?

I want to say something about the difference between big models and big data base estimations. I, myself, am not too cheered up by big data base estimations. But I think that big models are a fabulous substitution of capital for labor. The reason is that intelligent aggregation that is appropriate for a specific problem is much more difficult to do than to deal directly with disaggregated data and disaggregated phenomenon.

Think about the problem of building a national one-region supply-demand model for coal as a simple model versus using a 50 or 100-region model which is complex. The multi-region model is big dimensionally but not complex because it is simple behaviorally. You can see the explicit workings out of the very important transportation phenomena which are, in fact, decisive in the economics of coal supply and demand.

So let me say again that people don't believe models. That is not the issue. People use models to organize the impression they have in their heads to get on with sorting out the consequences of decisions.

DR. MAYER: I want to say something, Saul, that is in David Freedman's defense.

We in the Analysis Center have looked very thoroughly at the national gas debate over deregulation that was referred to. We have a list of quotations from Federal public officials, including people from the Department of Energy, endorsing models as scientific apparati.

The fact of the matter is that these models are presented as the latest in scientific analysis, particularly to the public. Now the fact that you and I know better, David, particularly since you build them, doesn't alter the fact that they are presented that way; the public believes it, the New York Times believes it, the Atlantic Monthly believes it, the New Yorker believes it, congressional staffs believe it, or some congressional staffs. The claim for these as science goes on repeatedly, particularly when the heat is on.

DR. GLASSEY: Let me pick up on the last two points that have been made here. David Nissen points out that models, if they are used by a decision maker, can be used by the decision maker almost in an interactive mode. And if the decision maker doesn't like the negative .60 elasticity, he can put in zero. He believes that more and then can use the model for designing his policy.

I don't think that anybody would argue about that particular use of models. I think where we get in trouble is when, as Larry said, the decisionmaker says, "I like zero better than minus .6" He then makes his decision and uses the computer printout that was produced with his explicit assumption of zero as allegedly scientific evidence to support the decision that he made.

At that point, the modeling community goes along with the story. I think we are not making the sort of contribution or rational debate that we should.

DR. MOSES: Well I wondered if any member of the panel thought that modeling was giving science a bad name.

DR. GASS: We will have Harvey's comment and then we will have one last question from George.

DR. GREENBERG: I wanted to briefly comment on this exchange. I wanted to state, in another way what I think is agreement with Dave Nissen, and that is that I think the statistical perception and approach to this kind of thing is that statistics is mostly about the data and that modeling is mostly about the relationships among the data and what they represent. The way in which models are used is to understand the relations and the implied relations that are not well understood at the outset, but get understood during the analysis process.

With reference to Dave Nissen's presentation on the electric utility example, if you took the simpleminded approach that I think would have been recommended by David Freedman in 1975, you would have used the historic growth rate of 7.2 percent, which is what was being used to justify the $100 billion subsidy by the Federal Government. But the modeling--despite its inequities--was still able to surface some of the understandings that led to the lower growth rate projection.

DR. GASS: One last question. George, did you want to comment or do you have a question?

DR. LADY: I agree with David Nissen, so I can ask him the hard questions. Many of the remarks that have been made about why it is all right, even if the data are no good or other things are no good, seem to me to refer to "modeling as an enterprise" which is conducted in a close staff relationship to a decision maker.

There has been a fair amount of emphasis of having the use of the model well understood and having the fact that we are supporting decisions willy-nilly of the apparent inaccuracies and so forth.

It is my insight that now, if not always, that most of the resources that have been spent have not been spent in that mode. And, indeed, for applied analysis nowadays we are producing a report that uses more than half of the resources on an annual basis and which presumably is being established as an information product report that stands as a grandchild of the Project Independence Report.

I don't understand it. It is a legitimate inability of me to understand how that information product can be offered without the degree to which it can be believed, somehow understood, and communicated because the decision maker that would be the recipient of the report is unknown and faceless in general.

I just don't know how we can--I agree with you that you have to go on, I understand all of the advantages of bringing the discipline of the modeling to the data that we have, but I don't understand how we can say that believing or understanding the accuracy can be ignored.

DR. NISSEN: No, I didn't mean to say that, George.

DR. LADY: That is what you said the data would be.

MR. NISSEN: Let me restate what I said. In fact, I have elsewhere argued, or argued previously here a year ago, that the modeling process had been used in mystical ways to bully people; that the emphasis on assessment within an independent EIA was an institutional response by the Congress to precisely that class of bullying and that assessment was stipulated and wisely, productively stipulated to be part of the modeling process.

I don't object to more assessment. I don't object to good science. In fact, I do some on occasion. But what I object to is the false syllogism that if the science isn't very good, then the modeling effort isn't worthwhile.

So I think that the assessment process ought to report the state of the science.

DR. LADY: Oh, I agree entirely. I said I agree with you. Yes, absolutely.

DR. GASS: With that happy note of agreement which I didn't think we would get to, I would like to call this afternoon's session to a close.

A Note on the Human Factors Associated with Model Validation and Confidence

David A. Pilati
National Center for Analysis of Energy Systems
Brookhaven National Laboratory
Upton, New York

As a participant in the NBS Model Validation Workshop, I felt a great deal of dissonance about its approach and neglect for what I would term the human factors associated with modeling and beliefs in model confidence. Our cultural tendency to objectify all endeavors was certainly in evidence at the NBS meeting. However, issues related to the more subjective nature of reality were raised on occasion but skirted and not seen as primary issues. Perhaps a similar conference in the future could explicitly treat these issues, however vaguely treated here.

On one level it seems that the Workshop's goals are to "scientifically" validate a process (i.e., energy modeling) that may have not been undertaken as a scientific endeavor in the first place. For example, the discussions on model validation raised the question of whether science or pragmatism rules in model development. I concluded that with increased demands for model results and limited resources pragmatism will govern model development more than science. If this accusation were true, how does one justify a scientific validation of such a process? This paradox lead one prominent economic modeler at the workshop to state that "validators are whiners." If we are indeed trying to scientifically validate non-scientific modeling efforts, this perception should be expected.

In the discussion on model confidence the human factor was raised again. Figure 1 is a schematic displaying a computer model and various levels of human interaction with that model. Conceptually, any model has an objectively measurable degree of confidence that could be quantified in some appropriate manner. However, this is rarely done in practice and confidence in a model stems from an individual's perception of the model based mostly on interactions with other model users and/or developers. These individuals are associated with institutions that color their perceptions and responses concerning their confidence in a particular model. When an individual is trying to promote the use of a model, how does their expressed confidence differ from what they actually believe? How does belonging to a particular institution color a person's perception or responses? For example, the model developers (inner ring of Figure 1) are usually aware of many model limitations or distortions that never are transmitted to other users. Model developers generally promote their products by emphasizing the positive aspects of their approach rather than its negative aspects.

Future workshops on model validation might explicitly consider the more subjective aspects of this topic. What are the institutional forces acting on the individual to decrease his/her honesty in appraising the quality of a model? How much of an individual's confidence in a model is obtained from documentation, personal use of the model and word-of-mouth assessments? What are the institutional pressures on a model developer or user that result in distortions of a model's appraisal? How can a model user correct for these biases in assessing a particular model?

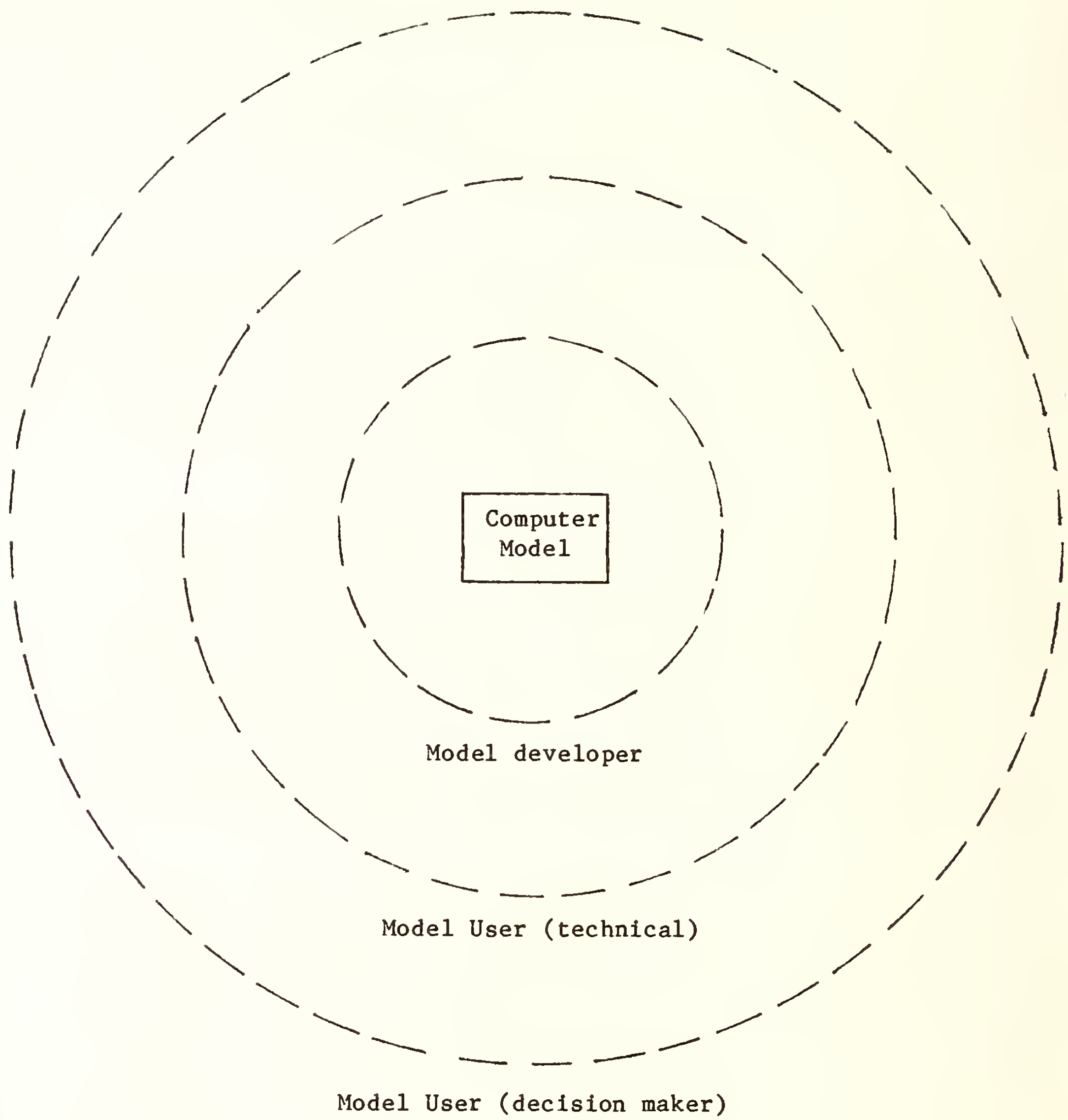


Figure 1. Levels of Human Interactions with Models

REPORT ON WORKSHOP 1

COMPOSITE MODEL VALIDATION

ANDY S. KYDES* AND LEWIS RUBIN**

Abstract

This paper summarizes the ideas and conclusions reached by the participants of the National Bureau of Standards workshop on Composite Model Validation held in May 1980. The focus of the discussion centers around the question: "Do composite models cause unique problems in achieving the objectives of the assessment process?" The participants included in the workshop were:

Robert Aster, Jet Propulsion Laboratory;
Neil Gamson, Energy Information Administration (EIA);
Rodney Green, EIA;
David Hack, Library of Congress;
Malcolm Handte, Wharton Analysis Center;
Andy S. Kydes, Brookhaven National Laboratory, co-chairman;
Norman Miller, Federal Energy Regulatory Commission;
Lewis Rubin, Electric Power Research Institute, co-chairman;
Jeff Sohl, Largo, Maryland;
David Strom, Conservation Foundation;
Chuck Weisbin, Oak Ridge National Laboratory; and
David O. Wood, Massachusets Institute of Technology.

I. Introduction

The intent of this paper is to focus on unique problems and issues related to composite model validation. This working group discussion focuses on the distinction between composite and non-composite models and the unique problems related to validation which composite models may introduce.

Methodologies developed for energy systems analysis have continued to evolve toward greater complexity and detail, reflecting in part the greater sophistication and understanding of the energy planning and analysis community. Since energy policies can ultimately affect the future patterns of energy production and utilization, planning, analysis and evaluation activities must be executed in sufficient detail to exhibit the effects of these changes on the supply and end-use sides of the corresponding energy systems and on the economic and environmental systems of which they are components. Further, choices among various policies and programs must be analyzed in the context of social, economic, environmental, and security objectives so that a balanced and diverse set of options may be identified and implemented. The multi-dimensionality of the planning objectives governing these choices has necessitated the integration of detailed process/product-oriented energy and environmental systems models and traditional stock-flow economic models. These integrated methodologies, which loosely speaking are composite models, have introduced a higher level of complexity for model assessment since the information flow within the model structures is usually more difficult to trace.

The major issues debated in this workshop were:

- What is a composite model?
- What are the objectives of model assessment?
- How do composite models create unique problems for assessment?

*Brookhaven National Laboratory
Upton, New York

**Electric Power Research Institute
Palo Alto, California

The need to develop an appropriate research agenda for composite model validation was also identified and remains an open issue.

Section II provides a working definition of a composite model and a working hypothesis to be proven or disproven. Section III discusses the objectives of assessment and Section IV deals with the unique assessment problems associated with composite models.

II. What is a Composite Model?

Any discussion focusing on the unique characteristics of composite model validation must presuppose unique properties of the composite model. The first step in the discussion, therefore, is to decide upon the boundaries. Where do composite models begin? Where do they end? What makes them unique?

In the course of ongoing debate, numerous attempts to answer these questions have been advanced. These can be summarized in three major lines of argument:

o A composite model involves the flow of information in more than one direction. More specifically, it involves the simultaneous determination of prices and quantities, each depending on the other. As an example, note that all integrated models of supply and demand, such as the Long-Range Energy Analysis Program (LEAP) of the Energy Information Administration (EIA), are composite in this sense. The Sweeney Auto Model, developed for the Federal Energy Administration (FEA), is not by this definition.

o A composite model is a hybrid, constructed by joining a number of other, less comprehensive models which describe subsets of the hybrid's universe. Independent development of the sub-models is usual but not necessary. Examples include the Mid-Term Energy Forecasting System (MEFS) of the EIA or the TESOM/LITM models of Brookhaven National Laboratory (BNL) and Dale Jorgenson Associates (DJA). The ICF coal and electric utility model, however, is not composite under this definition.

o A composite model is any model which explicitly represents more than one distinct activity. This notion is conceptually broader than the previous two in that it need not involve model linking, simultaneity, or independent development. It addresses the difference between a single-equation, price-quantity description of gasoline demand and a multi-equation, stock-adjustment, efficiency-weighted approach. Examples of composite models under this definition would include the Sweeney Auto Model and the Hirst Residential Energy Use Model of Oak Ridge National Laboratory (ORNL). Very early energy demand models, such as the FEA's short-term petroleum demand model, are not composite under this criterion.

Clearly, these arguments are not all-encompassing nor are they mutually exclusive. Further, it is not clear from the debate that an absolute statement is necessary or even helpful. A working definition, however, was adopted for the purpose of advancing the discussion. It is stated as follows:

Working Definition - A composite model, for the purposes of this discussion, is considered to be any model which is:

a) made up of separate components, independently developed, which were not originally designed to be compatible, and;
b) built by integrating (i.e. linking) two or more separate, dissimilar types of methodologies.

To help focus and discipline subsequent discussion we incorporate the definition in a working hypothesis:

"Mixed methodologies and overlapping components do not cause unique problems in reaching the objectives of the assessment process."

Before proceeding with a discussion of whether or not the hypothesis is acceptable under the working definition of a composite model, it is necessary to also characterize the objectives of the assessment process. This discussion is presented in the next section.

III. The Objectives of Assessment

Three broad objectives of assessment are clearly identifiable for all models:

- o to determine the appropriate uses of the modeling system;
- o to quantitatively estimate uncertainties in model results when the model is used appropriately; and
- o to communicate the information from the first two items to users outside the immediate analytical circle.

The first objective identifies the appropriate uses of a particular model by analyzing the problems and questions it was designed to address. Models cannot be analyzed before their appropriate uses are identified. Further, they should be judged by what they were designed to do. Misuse and misapplication of models is usually an error in judgement by the users; it may also reflect poor model documentation.

The second objective attempts to estimate quantitatively the conditional uncertainties that are associated with the results of a particular model. There are two aspects to the second objective. The first includes the scientific aspects which are often characterized by the more formal statistical procedures. The second includes the nonscientific or judgmental approaches. Although the uses and questions addressable by models are broadly defined and non-technical in nature, the analysis portion of the assessment process is more narrowly and technically defined[1,2,3,4,5]. The assessment of energy policy models seems to require a combination of both scientific and judgemental aspects. The appropriate mix is unknown, however, and does depend on the particular model. In this sense, model assessment is both art and science.

The third objective requires the dissemination of the information beyond the scientific community to potential users in a less technical but more

communicative way. The communication, to be effective, must occur in the natural language of the users. Two questions which must be answered are: "Does the model address user needs?" and "What is the importance of uncertainties in data input or model structure on the designated uses of the model?" In addition, suggestions for model improvements useful for a particular class of users or questions are important to communicate. Figure 1 illustrates the three phases of the assessment process.

Since the words "users of energy policy models" are often used, it may be helpful to identify them. Table 1 provides a comprehensive list.

Having categorized the assessment process, it is now necessary to examine the original hypothesis as applied to each of these categories. A judgement can then be made about the unique problems associated with the composite model in each phase of the assessment. The next section discusses these issues.

IV. Unique Assessment Problems Created by Composite Models

The working hypothesis of this discussion is that composite models <u>do not</u> create unique problems for assessment. As the previous section has illustrated, the assessment process has three distinctively different elements; thus it is important to examine each element in turn in order to decide whether or not to accept the hypothesis.

The first phase of the assessment process requires education of the assessor. It involves determining the intended functions of the model.

Discussions have indicated clearly that composite models pose no special problems for this phase of the assessment. Although no specific reasons were advanced to explain this, one possibility does come to mind. As stated in the working definition, the composite model is characterized by structural and methodological incompatibilities, both of which are technical or even mechanical in nature. The <u>concept</u> of the model, the <u>story</u> it tells, is not affected; and it is generally the concept - not the mechanics - which determines the function.

With regard to the second phase of the assessment process, however, the technical aspect is everything. The objective of this phase is specifically defined as the quantitative estimation of uncertainties associated with model results. The measurement and interpretation of such uncertainties are strongly affected by the mechanics of model linking; thus composite models may pose special problems here.

Discussions indicate, for example, that there are interpretive difficulties in combining uncertainties from a linear programming component and an econometric component in a composite model. Should they simply be added? Why? What other algorithm can be justified? Modelers and assessors do not yet know the answers to such questions, but nevertheless agree that the problems are a particular outgrowth of model linking considerations.

Similar problems arise in adding uncertainties generated by models outside the model of primary interest. For example, most energy demand models are driven by models of economic activity; thus part of the uncertainty associated with energy demand forecasts can be attributed to uncertainty in GNP

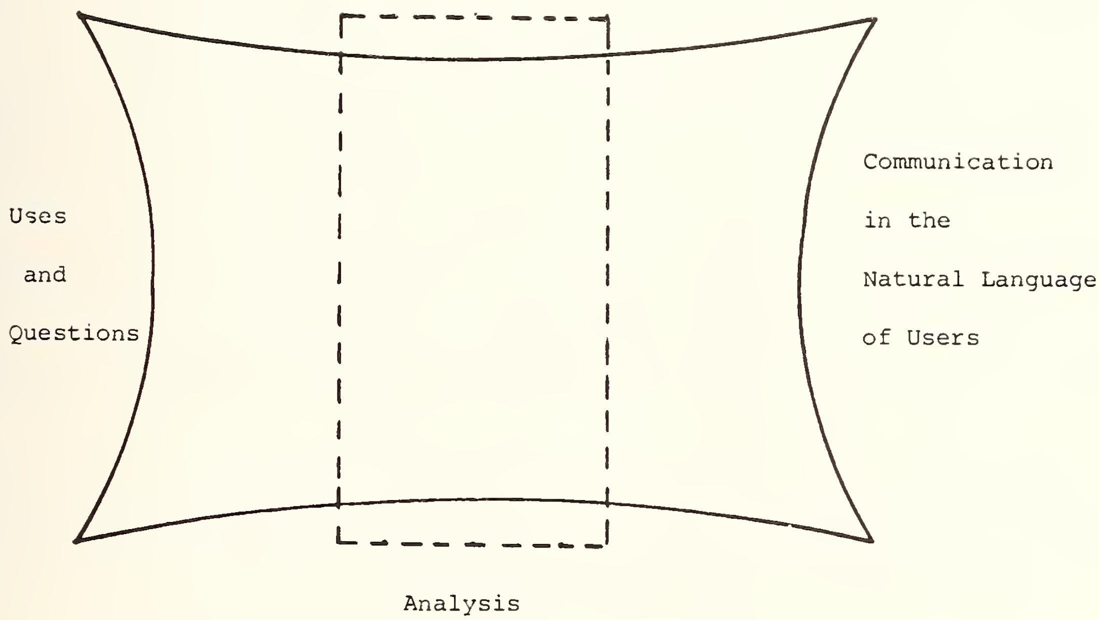


Figure 1: The "bare bones" of the assessment process.

MODEL USERS
Modelers
Peer Community
Model Sponsor
Assessment Sponsor
Model User
Analyst
Operator of Software
Other Analysts who Use or Interpret the Results
Decision Maker
Advisor to the Decision Maker
Constituencies Influenced by Model Based Analysis

Table 1: List of Potential Users of Energy Policy Models

estimates. This represents a linked system but one about which the assessor may have no information because the model of economic activity is outside the scope of his work. In general, the problem of measuring uncertainty in input variables which have been generated by other models outside of the current mandate is a composite model phenomenon.

Even in the absence of interpretive difficulty, however, administrative problems can confound the effort. If components of a composite model have been developed by different modelers, the assessor is forced to work with all of them in order to understand the entire model. This puts added burdens of coordination and interaction on the assessor.

As mentioned previously, limits in the scope of the assessment process also create problems for assessors. Such limits proscribe the assessor's ability to uncover and quantify all contributing influences to uncertainty of results. Further the proprietary nature of many component sub-models effectively limits the scope of the assessment in similar ways.

The third phase of the assessment process involves communication of the insights gained to the general population of model users. Discussions here indicate that composite models perhaps pose more difficult problems in that they themselves are more complex.

The linkages among submodels create an interactive representation of the world, which is perhaps more accurate but which also makes communication of concept and insights more difficult for assessors. It is noted, however, that it is precisely the complexity, interaction and comprehensiveness of composite models which makes them so attractive to policy makers and users. It seems that the problems experienced in communicating composite model insights are formidable but the results are worth the effort.

In summary, there appears to be very little evidence that composite models create a unique <u>kind</u> of problem for the assessment process. In general the <u>degree</u> of difficulty is greater for a composite model assessment, but this is perfectly reasonable for a larger model. A recap of these conclusions is given in a more familiar model assessment context in the final section.

V. Summary and Conclusions

Three major conclusions were reached in the discussions which provided material for this paper:

- It is not at all clear where one should draw the line between composite and non-composite models. In fact, one would be hard-pressed to identify a non-composite model being used seriously in the energy analysis community today, because almost all models have elements of "compositeness" about them.

- With the exception of a very specific, albeit quite important, problem concerning the combining of uncertainty estimates, it appears that composite models pose no unique assessment difficulties. In general, the problems are more difficult but not different.

- In general, composite models are the preferred tools of most model users, because of their comprehensiveness and feedback effects.

Table 2 illustrates the results of the hypothesis testing in terms of a somewhat different - and more familiar - disaggregation of the assessment process[6]. When viewed in this way, the acceptance of the hypothesis is clear.

Hypothesis: Mixed Methodologies and Overlapping Components <u>Do Not</u> Cause Unique Problems in Reaching the Objectives of the Assessment Process	
<u>Element of Assessment Process</u>	<u>Hypothesis</u>
Validation (third party responsibility)	
Structure and Specification	Accept (subject to the complication concerning the combining of uncertainties)
Data	Accept (subject to the complication concerning the combining of uncertainties)
Content	Accept
Prediction	?
Evaluation of Documentation	Accept
Evaluation of Useability	Accept
Verification (modeler responsibility)	
Documentation of Code	Accept
Impact Analysis of Code	Accept
Documentation of Impact Analysis	Accept

Table 2. A Test of the Composite Model Hypothesis

References

1. Fred Schweppe and James Gruhl, "Systematic Sensitivity Analysis Using Describing Functions", NBS Special Publication 569, January 10-11, 1979, pp. 497-515, Saul Gass, editor.

2. R.G. Alsmiller, et al., "Interim Report on Model Evaluation Methodology and the Evaluation of LEAP," Oak Ridge National Laboratory (ORNL/TM-7245), April, 1980.

3. Yvonne Draper, Andy S. Kydes, and Steve J. Finch, "A Statistical Approach to Model Simplification and Multiobjective Analysis", to Appear in <u>Energy</u>.

4. E. Kuh and D.O. Wood, <u>Independent Assessment of Energy Policy Models</u>, EPRI EA-1071, May 1979.

5. Martin Greenberger and Richard Richels, "Assessing Energy Policy Models: Current State and Future Directions", Annual Review of Energy, 1979, pp. 467-500.

6. Saul Gass. "Evaluation of Complex Models", <u>Computers and Operations Research</u>, Vol. 1, No. 4, March 1977.

REPORT ON WORKSHOP 2

THE MEASUREMENT OF MODEL CONFIDENCE

Lambert S. Joel*
John S. Maybee**

Although the workshop was nominally intended to explore methods for the measurement of confidence in models, the discussion ranged over various topics from defining "model confidence" through measuring it, to establishing and increasing it, with major emphasis on the last of these. The workshop's initial discussion paper (which follows this report) is substantially devoted to questions of confidence enhancement through conventionally regarded sound practices in model formulation and documentation. This approach, ignoring problems of definition and measurement, involves the reasonable (tacit) assumption that confidence is directly related to quality. A consensus identified six areas suitable as reference points for consideration of questions of confidence. These six subjects comprise a checklist for confidence building, that intersects strongly with the independently derived ideas in the discussion paper.

Prior to elaborating the checklist, we review the workshop discussion of the nature of models and confidence in models that established a background for joint adoption of the confidence list.

(1) Confidence can be defined informally as the belief (or degree of belief) that a given model can furnish information that will be useful in solving some problem. Rather than attempting to make this simple notion precise, we can give a few recognizably common examples to clarify the decision context: willingness of managers to condition courses of actions on weather forecasts known to be based on large scale models; reliance on statistical demand forecast models in marketing and production planning; and (confident) use of instrument navigation and landing aids by commercial airline pilots.

(2) Confidence is affected by the experiences and prejudices of prospective users of models and model results, and people have individual standards for acceptance. Moreover, acceptance sometimes seems to reflect preferences rather than judgment: almost everybody will always place much greater reliance on model outputs that support their opinions than on those which tend to refute them.

(3) The role of understanding is not clear relative to confidence. The "public" doesn't understand models in general and is skeptical of model results. In contrast, there is rather universal acceptance of sampling and estimates acknowledged to come from sampling, but the degree of public understanding of sampling methodology is no greater than the understanding of models in general.

----

*National Bureau of Standards, Washington, DC.
**University of Colorado, Boulder, Colorado

(4) It can be argued that models can not be judged objectively by policy makers who are too close to the decision problems. But it must be noted that "conventional" assessments by professional analysts are usually couched in narrative terms, and that apparently mechanistic methods of evaluation of model characteristics and model outputs are subject to sufficient leeway in execution and interpretation that the differences in subjectivity of treatment between analysts on the one hand and everyone else on the other may not be very great.

(5) "Dumb luck" plays a role in model acceptance. "Good looking" intermediate outputs can promote a bad model; a good model with a single conspicuously false assumption might be rejected.

(6) If a model has been in existence long enough to have been applied several times, the "track record" might furnish a basis for determination of confidence. If data are very scanty or uncertain, the quality of previous model results may not be significant. The existence of a large number of models for which details of structure and outputs are not publicized--those produced within corporations for internal application--might seriously distort any judgment of collective track record of models as an indicator of state-of-the-art.

(7) There is widespread resistance in the "user community" to modification of models in use, irrespective of quality, possibly because previous applications are perceived as furnishing a basis for comparison of output. This implies that users would prefer to post-process, (massage), model outputs in the light of new information than to employ a revised model.

(8) An interesting "third-party" method for evaluation, or determining degree of confidence in a model if you have a lot of money: The Nuclear Regulatory Commission has on occasion parceled out to minor contractors portions of the analysis for which a large model had been developed. The contractors were required to generate independently solutions to these small problems. The results were then compared to the relevant big model outputs. In the instances described, agreement was said to be satisfactory.

(9) Parsimony is important, up to a point: models that are "too big" hinder confidence, while exceedingly simple models are dismissed frequently as being "back of the envelope" analysis,

(10) Concern was expressed for the lack of measures of model "confidence worthiness" similar to measures of forecast accuracy. Oddly, participants whose stated professional assignment is assessment said they had come to our workshop in search of approaches to assessment de-emphasizing quantitative methods, which had been unsatisfying. The role of measurement was discussed briefly. A claim that it relates only to validity "which differs from confidence" was countered by arguments that validity is a factor in confidence and that "speed of response," e.g., is quantitative, and affects confidence but not validity.

(11) The possibility of ordinal measures for acceptance was mentioned, but a statement that an ordinal approach might serve to compare models and not provide a means for acceptance or rejection of a model in isolation foreclosed further discussion.

(12) There was not universal agreement in the group about the degree of importance of tests. A claim that tests were an absolute prerequisite to establishment of model confidence was countered by an argument that sometimes tests can't be made (e.g., long range forecasts) and that tests themselves can be defective. (Note: Clearly both viewpoints are valid with some hedging. This matter was not pursued further, but as can be seen below, testing is an entry on the consensus checklist).

This is a committee report. The foregoing material is a freely edited transcript of ideas that were set forth and discussed by the session participants. What follows here, however, represents consensus that there are at least six matters of great importance to the establishment of model confidence, i.e., six areas in which one could profitably investigate questions of confidence. What results is a sort of checklist that can be of benefit to model makers, model users, or model assessors. It resembles, in fact, a typical assessment checklist, but as will be clear in the elaboration of individual components, it diverges somewhat from conventional assessment orientation.

The six confidence related areas are:

- Algorithmic Integrity
- Software Quality
- Quality of Data
- Benchmarking
- Scenario Support
- Model Management/Data Base Administration

No importance ranking is attached to positions in this list. The notion that some attributes may affect confidence more than others was not discussed in the one-day session. The labels for the six subject areas were assigned for convenience. This extended descriptions below will clarify the intended meaning of each.

Algorithmic Integrity: Refers to how input is to be transformed into output, i.e., the conceptual methodologies of the model and their mathematization, or the algorithms of the transformation. Whatever methods are used, e.g. exploratory curve fitting, extrapolation by time series analysis, static equilibrium descriptions, or detailed dynamic process simulations, several natural confidence oriented questions arise: Why was the method chosen? How is it situated in terms of state-of-the-art? Under these headings are a host of issues related to mathematics, statistics, numerical analysis, and systems analysis; robustness, flexibility, breadth of focus, numerical stability, etc. which can be addressed in informal as well as technical terminology.

Software Quality: Involves numerical precision and reliability (an area of overlap with algorithmic integrity), as well as portability of computer models, the extent of built-in facility for modifications, and the convenience of input/output procedures.

Quality of Data: All of the data for a model, the nominal inputs for model runs, as well as data for determination of the values of model parameters, can affect the performance and utility of a model. A brief catalogue of quality factors includes identification of sources, relative completeness, uniformity of precision and procedures for accomodating outliers and missing values.

Benchmarking: This term is used in computer software management to denote test computations made for reference purposes. Here we include all sorts of output recording and/or testing procedures, a few of which are: (1) Sensitivity analysis and stress testing (by abstract mathematical methods as well as computationally); (2) Freezing a model with associated data at various stages of development; (3) Comparison of ex ante and ex post forecasts. Sensitivity analysis measures the reponse of a model to changes in data and parameter values, or computational methods. Stress testing probes the "elastic limits" of a model by sensitivity analysis using extreme values. Sensitivity analysis of big models is tricky because the straightforward method of varying one parameter, say, at a time while holding everything else constant, (analogous to partial differentiation) sometimes fails to yield enough pertinent conformation. Then, sophisticated methods of selecting the "right" groups of variables and parameters must be used in order to avoid the computational expense of testing all possible combinations. Ex post and ex ante are terms used in econometrics. Ex ante means that I use today's model to forecast, say, one year into the future. Ex post, sometimes called "back-casting" means that at the end of the year I correct the parameters of the model according to the end of the year data and then "forecast." Of course, the dependent variable cannot be used as a datum for correcting the values of parameters, whereas in constructing the model originally, all data, including available time series of observed values for the dependent variable, could have been used in setting parameter values. Such a comparison helps to identify the relative effect of the structure of the model and the adequacy of support data on the accuracy of the forecast.

Scenario Support: Several ideas form this area. The first is extraction of intermediates. This idea is related to the following question: "If a model gives results which are unexpected or counterintuitive, you would like to be able to explain why this happened?" Presumably your ability to recover intermediate steps will help to do this. Organization of a model that would allow you to extract intermediate results would be extremely useful and certainly can be expected to enhance confidence in the model. Next, we would

like to know really, for what sets of inputs you can expect the model to give you worthwhile results. Information about this sort of thing is clearly confidence related particularly for the users of models. There is some overlap here with Benchmarking and with Quality of Data. The frame of reference under those headings, however, was basically undifferentiated, "abstract" questions of accuracy and model system characteristics; here the emphasis is on the purposes of the model. Another subhead under scenario support is baseline and alternative scenarios. The principal notion here is that of analysis and explanation. Thus, we are interested in transparency in formulation that will permit convenient correspondence between narrative description of scenarios and the sets of parameter values which define the scenarios formally, (i.e., mathematical "state space" specifications). Assessors (and users) will want to run a variety of scenarios and be able to tell a good story about the outputs.

Model Management/Data Base Administration: This refers principally to the environment in which a model is operated. If a model is to be used on a daily basis, or at least, "frequently", the operational environment is extremely important. The extent to which this environment is systematic and orderly has strong confidence implications. Model maintenance is a critical component of model management for models amenable to continued application. It involves planning for archival procedures and updating of all components of a model and associated data, as the analysis of a subject system evolves in response to "exogenous" information growth and feedback from prior operation of the model. An associated issue in management/maintenance is the establishment of a directory of who knows what, or who has been responsible for what, during model development or operation. This sort of information facilitates modification of a model by persons other than the original developers, and provides a means in general for redressing lacunae that always exist in documentation of a complex model. Directory information may be embedded in a history of the model, a category of information of perceived importance roughly related to the frequency, duration and range of actual applications. (That is, for complex models in heavy use, a history is more readily acknowledged as an aid in interpretation of model results, while a history for record purposes of a once or twice-employed model may be considered merely cosmetic). A directory or a history are instances of a central confidence-related notion; that of an audit trail, which indicates the provenance of various entities of the model. It is especially useful as a trace of the sources and organization of data, and this tells part of the story of how the model results came about.

Documentation, a general confidence related area for models, has not been accorded a separate listing here. Its importance can be gauged by noting that it is threaded, at least tacitly, through all six of the items in the checklist, and invoked under several aliases in the sixth. The session heard a brief statement on a significant intersection between computers and documentation. Dr. Harvey Greenberg of DOE cited work under way in OAOA aimed at automating the production of portions of audit trails and other aspects of model documentation.

We close this account with some remarks by the session co-chairman Dr. John Maybee given at the end of his oral report of the session discussion: "This, I think, summarizes the things we have to say, that I want to report about. I want to exercise the chairman's prerogative at this point, and make one or two comments of my own feeling about what was done in our workshop at least."

"We did a lot of talking and we came up with a list of things which are I think fairly obvious things, but I suspect that of the people in our workshop, there was really only one present who has done any serious assessment of models. It seems to me that what we've done, up to this point, is probably useful, but we've been talking about all of this for a year or two; now it's time to get down and do some of it. We must do some hard analysis on some models and quit talking so much, because we can make these nice lists, and we can enhance these lists--I'm sure we could flesh out what was on that chart (the checklist) to a considerable extent, but I really feel that the real challenge at this point is to get busy and make precise and apply some of these slippery philosophical notions. I hope that if and when there's another symposium or conference of this type that we will have some reports and papers of assessing and confidence building actually addressed to specific models and we'll see something more than just talk about how we think it might be done."

MODEL CONFIDENCE DISCUSSION PAPER

We begin with five postulates cast firmly in sand:

(1) Confidence in a model is subjective, in contrast to its major ingredient--validity;

(2) Validity is objective, but context dependent;

(3) For the analyst, confidence in and validity of a model are so intimately intertwined (read correlated) that they are not worth distinguishing;

(4) For the non-analyst model user, smart but no technician (read mathematician/statistician), confidence in a model comes from a variety of factors beside validity (which has a different meaning than it has for an analyst, anyway). User confidence is a core subject of this position paper;

(5) Analysts can enhance "legitimate" confidence for users only by exercise of a model and lucid exposition of model analysis to a "user community."

The remainder of our paper will elaborate these assertions:

(1) Is merely a tautology as stated, but its importance here is that we will not attempt to define nor seek measurements of "confidence-worthiness" as an objective attribute of a model. We feel that a talented (sophisticated) user will usually have more confidence in the employment of a model than an untalented user; but, on the other hand, an untalented user may respond to indicators of model worth that can be spurious, such as the presence of a large population of his peers who are "satisfied customers;"

(2) Although no precise definition of model validity has yet appeared, we find it convenient to consider all those attributes of a model and its performance which are measurable (if only in principle), as the constituents of validity. Although many of these will depend on the purposes of a model's employment and the milieu in which that occurs, they are not transformed by the personality or attitudes of the user. Hence, the notion of context dependence and the distinction between "subjective" confidence and "objective" validity;

(3) Is a moral ideal, reflecting the belief that a "compleat analyst" is totally free from bias or emotional commitment to (or against) any model, so that subjective confidence does not obtrude onto the analyst's evaluations. We thus dismiss further consideration of analyst confidence;

(4) Reasserts the idea that the user (policy maker, e.g.) responds to a variety of signals relating to how a model has satisfied or appears likely to satisfy information requirements for his purposes;

(5) We can't see any way to tell a user how confident he should be of a model through bare recitals of our objective criteria.

Ultimately, the utility of a particular modeling structure or algorithm depends upon the insight and understanding of the user. Thus the utility of a model is a subjective matter and may be expected to vary from one user to another. We will devote the sequel to aids for disseminating insight. Some will object to the above assertions on the grounds that a model designed to make forecasts can be compared with reality and a "track record" established. This may in fact be true for short-range forecasting, but for longer-range forecasting the utility of a model as a predictor may be understood in a different sense. In the long run, its future is not necessarily likely to closely resemble the past. In fact, policy decisions based upon forecasts made using mid- and long-range models may act to insure that the future is quite different from the past. Thus, such models may prove extremely useful even in situations where their track record would turn out to be very poor. The fact of the matter is that such longer-range forecasts often only need to be qualitatively correct. Policy decisions are apt to be based upon the following types of information:

(1) An increase in x leads to a decrease (an increase) in y;

(2) If both x and y decrease, then z will also decrease;

(3) A large increase in x leads to at least so much of a decrease in y.

The statements labeled 1 and 2 are purely qualitative in character. Usually, we expect our models to provide more than just such simple-minded qualitative information. The statement 3 is an example of such a more robust piece of information. It thus represents the kind of forecast we might expect to obtain from longer-range models.

We do not wish our point here to be misunderstood. Nearly all models actually lead to quantitative forecasts, i.e., to real numbers. The users must understand how to translate a set of such quantitative forecasts into statements such as 3. This is one of the ideas we had in mind above when we asserted that the utility of a model depends upon the skill of the user.

If we are correct in our assertion that confidence in a model is subjective, then it seems likely that the way to transmit confidence is through analysis. Therefore, let us turn next to our second basic position.

We shall take the fundamental position that analysis of a model consists of documentation of the model together with the application of any procedure designed to evaluate some aspect of the utility of a model or to detect a deficiency in it.

A careful description of this position would begin by separating the analysis methodology into several distinct categories and proceed with a lengthy and detailed investigation of each such category. But this is a position paper designed to provoke discussion so we shall keep it short and simply enumerate in shotgun fashion a list of points. Our hope is that participants will bring up additional points and/or show us where we are incorrect.

(i) One must have a description of the assumptions made by the modeler, with reference to the "world" that is modeled. One must try to decide whether or not the model is plausible and to what extent it has experimental support.

(ii) A description of the modeling methodology used in the formulation of the mathematical model is required. It is important to try to decide if the choice of the methodology is adequately justified. Also we must know to what extent the structure and equations of the mathematical model are consistent with the theory and methodology.

(iii) A logical decomposition of the model into convenient and recognizable submodels with flow charts indicating clearly the variables of the model (input variables and output variables) and the parameters is required. We must try to determine whether or not the parameter values set by the modeler are plausible.

(iv) For each submodel, a list of the equations needed to produce the output vector $\bar{y}$ given the point vector $\bar{x}$ is required.

(v) For each submodel and also for the entire model, a definition of the domain, i.e., the region in $\bar{x}$-space for which the model is claimed to perform satisfactorily, must be stated. We must then ask, if when the output from one submodel acts as the input for another, will the former always be inside the domain of the latter? We must also ask, does a unique solution to the equations in each submodel and in the entire model exist for every $\bar{x}$ in the domain?

(vi) For the parameters fixed by the modeler, we must know both their nominal values and some quantification of uncertainty (preferably a probability distribution or some properties of it).

(vii) We must have a statement concerning the existence and uniqueness of the solutions to the equations used in each submodel and in the entire model.

(viii) We must have descriptions of, or references to documentation for, the algorithms used to solve the model equations, and a statement regarding the convergence properties of these algorithms. We must know if "state of the art" algorithms have been used to solve the model equations. We must know if there are portions of the programs that could be designed better in the sense of reducing roundoff error.

(ix) We require a quantitative description of the approximation (convergence rules, mesh sizes, discretizations) built into the algorithms.

(x) We must have a documented listing of the computer program and the instructions for its use.

(xi) We require a description of all parameter estimation procedures, including (a) the way the data were obtained and (b) the method of estimation including both the statistical and numerical assumptions underlying the method. We must ask, are the uncertainties given for parameters that were formally estimated consistent with the statistical results of the estimation procedure? Are the parameter estimation procedures and the demonstration of goodness-of-fit statistically valid under the assumptions given?

(xii) We must have a description of real world data presented in support of the model including (a) how the data were obtained, (b) quantification of uncertainties in the data in probabilistic form, (c) the settings of the x-variables and the parameters used in computing the output, (d) a description of any formal statistical goodness-of-fit tests done together with the underlying assumptions of the tests, and (e) the role of these data in the development of the model.

(xiii) We must attempt to analyze the structure of large models and discover methods for simplifying and/or approximating them. This should be done on the theory that "small is beautiful" if it will work nearly as well as large. Qualitative analysis should be applied wherever possible. Experiments should be made with model output applying the principles of response surface methodology.

(xiv) Discretization techniques used to generate algorithms should be reviewed and the size of all truncation errors identified. The interplay between roundoff errors and discretization errors should be analyzed when appropriate with a view toward optimizing the total error. Stability problems, both mathematical and numerical, should be identified if possible. Alternative methods must be investigated where lack of stability occurs or is suspected. Sensitivity issues should also be studied.

(xv) Where possible, such techniques as Monte Carlo methods, stratified sampling, latin hypercube sampling, etc. should be applied to the analysis of uncertainties in models. An attempt should be made to fomulate general methods for assessing uncertainties in model outputs in terms of known or assumed uncertainties in model inputs and parameters.

(xvi) Attention should be given to the performance testing and fine tuning of any algorithms used in the model. Questions of proper data base management for the regular use of the model and to answer queries regarding the model should be answered and improvements made where appropriate. If state of the art algorithms have not been used and do in fact exist, they should be brought in-house and used to replace the algorithms in the model.

Now let us explain how we can use the above list of activities as a prescription for enhancing confidence. In the first place, not all of the above activities will be appropriate to a given model so that a sublist should be devised in each case. Then a small group of experts can be assembled and assigned the task of analyzing the model. This group would then report back their findings. How would such a report be useful?

Two groups of people might be identified as being concerned with confidence issues. First we have other potential users of the model. Presumably, they will be expert enough to understand the significance of some parts of the report made by the analyzing task force. Their level of confidence may then be enhanced to the extent that they can understand the report and to the extent that the report is itself satisfactory.

A second group of people concerned with confidence issues consists of policy makers who rely upon the output of a model to help them make policy decisions. The level of confidence such people have in the model must be a derived level of confidence dependent upon their confidence in the scientific ability of the analyzing task force. The seems to imply that the report should contain an executive summary in which the significance of the findings is explained in layman's language.

REPORT ON WORKSHOP 3

MODEL STANDARDS TO AID ASSESSMENT

DR. MILTON L. HOLLOWAY
EXECUTIVE DIRECTOR
TEXAS ENERGY AND NATURAL RESOURCES ADVISORY COUNCIL

In our workshop we were concerned with the standards that aid assessment. We built our discussion loosely around questions raised in the prepared issue paper. I wrote the issue paper and it reflects the major conclusions of a year-long assessment of EIA's Midterm Energy Forecasting System by a team in Texas which I directed.[1] The issue paper makes three major points. The first is that we should recognize that there are three basic kinds of models, and that when we talk about assessment, and about standards or guidelines to aid assessment, we need to recognize which kind of model is being assessed. The three types of models identified are disciplinary, subject matter, and problem solving models.

The second point made in the issue paper was that in this NBS conference and the one that preceded it, we were primarily concerned with the problem solving type model, and furthermore, that we are and have been concerned with large-scaled computer models and their use in the public policy process. My perception is that this conference, the model assessment activities at MIT, the Energy Modeling Forum, and the Texas National Energy Modeling Project are primarily responses to conditions that arose out of the use of large-scale models in the public energy policy process; therefore, we are and have been primarily concerned with that kind of model--large-scale models that are used in public policy decision making. In my classification system these are problem solving models.

The third point in the issue paper was that, given the problem solving, large-scale model orientation, the focus on assessment (and therefore on standards specifying to model builders what to provide to model assessors) needs to shift emphasis somewhat from the computer based model kinds of information to more information about the process--the way in which models are used in the public policy decision making process. This is because the process--the way problems are designed, the kinds of time frames that one has to operate in, the disciplinary backgrounds required by a particular problem--has a great deal to do with the choice of models, the choice of techniques, whether you build a model, or whether you use an existing one, and so on. The process needs assessment and models need to be assessed in the context of the decision making setting. This last point turned out to be highly controversial in our workshop.

We spent most of the day debating whether or not it is appropriate for assessment to get into the process, per se; I argued that is should. Some people felt that this was getting out of assessment of models into assessment of policy analysis, per se, and that that was a different kind of

---

1 Milton L. Holloway, ed., Texas National Energy Modeling Project: An Experience in Large-Scale Model Transfer and Evaluation, Academic Press, New York, 1980.

animal outside the scope of this conference. Others felt that assessing the process constitutes an evaluation of whether something called "due process" was really working right or not. And still others thought that to move in that direction was something that a political scientist might do, but that model assessors ought not to do.

So we did not reach any firm consensus on this point. However, there were several in the group who were supportive of including the process in assessment. The emphasis given in the issue paper was recognized as a needed part of assessment, but several of us came out of the workshop with different ideas of how strong the emphasis ought to be.

A set of charts helped focus the workshop discussion. If you ask the question, "What should be evaluated in assessment?" then my claim is that there are at least three classes of models--disciplinary, subject matter, and problem solving models--and that what is appropriate to assess in each of those cases is different. This outline is shown in Table 1.

Let us focus first on disciplinary models. Those are the kinds of models with which you and I are most familiar because the disciplines are where our formal training came from. Most of us come out of a particular discipline, and the process for assessment as practiced in graduate school programs and related professional societies is fairly well defined.

Disciplines are reasonably well self-policed so far as assessment is concerned, so I do not think this conference is very much concerned with strict disciplinary models.

The purpose of disciplinary models is to push back the frontiers of knowledge in a particular discipline. They are designed for educational purposes and so on, and not very much for problem solving.

Subject matter models (Table 1) include more than one discipline. They are models built around various subject matter areas--energy, water resources, population, so on. They collect techniques and methods--approaches from several disciplines--and focus on a subject matter area. But they are not problem specific. That is, they are designed to address a class of problems, not a specific problem.

The kinds of things that we worry about or ought to worry about,in my opinion, in subject matter models are pretty much the same as in disciplinary models. That is, we should examine primarily (1) the data, (2) the logic (or the theory), and (3) the actual behavior of the model in our assessment activities, but we should also pay some attention to comprehensiveness, because subject matter models have to address more questions than strict disciplinary models.

But I believe the primary concern of this conference should be with problem solving models, and therefore we have to look at a much broader range of issues in deciding <u>what to evaluate</u> and what kinds of standards people ought to adopt and use for purposes of aiding assessment. Heretofore, our concern has been primarily with the first three areas; the data, the logic, the behavior of the model--things that are tightly centered

TABLE 1

WHAT SHOULD BE EVALUATED DEPENDS*

Type Model	Data	Logic/ Theory	Behavior	Measure of Uncer-tainty	Compre-hensive-ness	Affect of Assess-ment	Process of Use
1. Disci-plinary	X	X	X				
2. Sub-ject matter	X	X	X		X		
3. Prob-lem Solving	X	X	X	X	X	X	X

Emphases of assessment needs to shift to these neglected areas and therefore information (guide-lines) to aid assessment must in-clude these items.

* The NBS Conferences are mostly concerned with standards for problem solving models since they (conferences) are EIA sponsored and since concern for assessment grew out of model use in public energy policy, and emphasis is on large-scale models.

around a computer code. The emphasis needs to move more to the latter four areas (Table 1), which are equally the concern of problem solving models used in the public policymaking process. These include some measure of certainty, which is more important in public policy and problem solving situations than it is in disciplinary models.

Comprehensiveness is a key measure of the adequacy of a problem solving model. If the model is poor in this way it is leaving out much information that the decision maker knows is important. Assessment of problem solving models ought to look closely at comprehensiveness.

We spent a good deal of time in the workshop discussing the effect of assessment itself as an assessment concern. That is, if we are going to be concerned with assessment, then one of the things we ought to pay attention to is the effect of assessment on the evolution of the model over time--this should be evaluated. As a model built for a particular purpose evolves--as they always do if they are successful--the question is how have assessments by various groups affected the changes made to the model to make it more useful?

Finally, the thing that generated the most controversy in our group, was the idea that assessment should include an examination of the decision making process and the use of the model in this decision making setting (Table 1). In the issue paper I claim that the NBS conferences--this year's and last year's--were mostly concerned with standards for problem solving models, since these conferences are EIA sponsored, and since the concern for assessment grew out of model use in public policymaking, and the emphasis has been on large-scale models.

The last column of Table 1 suggests that assessment should be concerned with the evaluation of the process in the case of problem solving models. Therefore we should ask model developers to supply a set of information related to the process in order to complete the assessment task; this information will be just as important as documentation of a mathematical model.

As shown in Figure 1 there are essentially three classes of participants in the public policy problem solving situation. First is the set of decision makers. Some people at the conference say they have never seen one of those characters. If that is the case, they are evidently not involved closely in policy work.

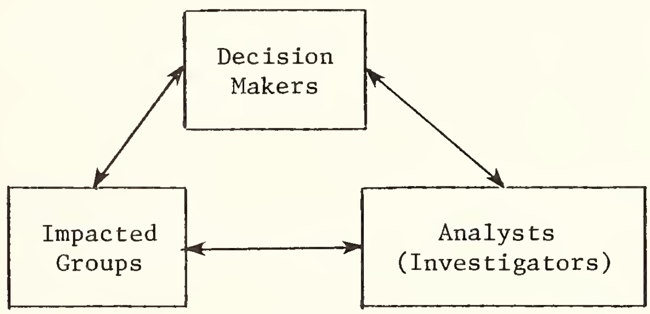
As the policy process operates in Texas, I certainly know who the decision makers are and I answer to them. And I believe the lines of authority are clearly defined in the power structure of government, whether it is in Austin or Washington. Although we may work for many clients, and there are many players in the final decision making process, those players are usually uniquely identified. They are primarily elected officials and appointees of elected officials; they are the center of power in the U.S. governmental decision making process.

A second group involved in the process is the analysts (Figure 1). Analysts are investigators, and most of the people in this conference fall into that category. Investigators include not only advisors and directors of research programs, but modelers themselves.

FIGURE 1

WHAT SHOULD BE EVALUATED IN
PUBLIC POLICY PROBLEM SOLVING MODELS?

I. Policy formulation in our political system involves three principal groups:



(To be responsive this group "banks" positive and normative information; models, books, data bases)

II. Assessment should include evaluation of how analysts behave and structure their work as a member of this process.

III. Process tells us something about needed model characteristics; this process puts a premium on:

A. Timeliness
B. Flexibility
C. Dynamics
D. Feedback

The third important group involved in the public policy decision making process is impacted groups--representatives of those people impacted by policy decisions. If you observe the process at work, this group certainly gets involved and decision makers always consult with them. And in my opinion, if we are going to do a good job as analysts in this setting, then we ought to interact with these people too, because such interaction is a great source of learning about which values different groups hold; that is, whether the various outcomes of policy decisions are going to be taken as good or bad by particular groups. And that has a lot to do with what is worth analyzing or modeling. Figure 1 indicates that there is, in the policy process, a three-way interaction between these three groups. Assessment should examine the adequacy of model use in this process.

The other point of Figure 1 is that in our behavior as analysts in this setting (in order to be responsive) we have learned to create "banks" of information. I mean that we create computer models and data bases, books and publications, including a set of normative information about good and bad outcomes that reflect the values of people impacted by decisions. Decision makers, especially elected officials, reflect these values if they get re-elected. Much of this positive and normative information can be banked, and in order to operate efficiently, that is what we do. A computer model that is the concern of this conference, for example, is a bank of information about relationships between important parts of the energy, economic and environmental systems.

To follow the argument a step further, assessment should include the evaluation of how analysts behave and structure their work as a member of this process (Figure 1). And my point is that one really cannot appropriately analyze a model--a computer model--and formalize assessment of models, unless the process is included. I will say more about the specifics of this part of assessment later in the paper.

The third major point of Figure 1 discussed in our workshop is that the policy process tells us something about desirable model characteristics. When we consider formalizing part of the process in the form of a computer model, then certain characteristics fit better than others. The first characteristic that should guide model selection is timeliness. In many situations, if you cannot get the answer by next week, or three months down the road, it is not worth getting. The political policy process puts a high premium on timeliness.

The second characteristic is flexibility. In order to be efficient one should structure and bank information so that it is flexible; one needs to use what is needed and leave the rest. Everyone knows that in a dynamic world everything is potentially related to everything else, implying a need for large integrated models. But at some point the value of impacts on remote parts of the system is very small. Models should be structured to allow flexibility.

The third characteristic is dynamics. All who have been involved in public policy know that the systems we try to model are dynamic, and that people are not worried about just today, but about tomorrow. Further, the decisions that they make take tomorrow into account. Models should capture time-dependent relationships. A policy solution involves "when" as well as "what" to do.

The fourth characteristic is feedback. In the political policy process a great deal of feedback occurs among the groups. There is feedback from

various places at various times. When we structure some of the relevant information in the form of a computer model, these feedback loops ought to be represented so that we can make changes appropriately.

The issue paper prepared for our workshop contained a checklist for model documentation. John Mcleod originally published this check list in Simulation Today. We did not spend as much time on this list as I intended but in general, I think people agree that Mcleod's list contains the right elements but there was considerable discussion about balance. Generally, there was agreement that Mcleod's list includes many items that a model developer ought to supply for purposes of an assessment. The outline of Mcleod's list is shown in Table 2. The first set is project information. This is very simple, straightforward information, but it is surprising how much of this is now included when one tries to assess someone else's model. The title for the model, the organization who sponsored it, a contact person, the objectives of the project, project duration, etc. should be included.

Second is model development information--the names of the modelers involved, purpose for which the model was developed, disciplines involved in the development of the model, the data required, methods of development, cost of development, availability, compatibility with other kinds of similar models, and the extent of use--should be made available.

In the workshop we did not think Mcleod's list contained two items everyone considers very important in examining and trying to understand a model. Explicit information should be given as to how the choice of parameters came about. Did the modeler use regression techniques, someone else's estimates, or some other procedure? Second, there needs to be an explicit equation specification. One cannot easily untangle a computer program without such information.

The third item on the checklist is model description. Classification of a model helps immediately to gain an understanding of the model. For example, one should describe whether a component is an econometric model, an LP model, a systems dynamics model or some other type.

Other needed model descriptions include a block diagram of how the model system works, the computer program itself, the notation used in the program, attempts to validate the model, reference information, the distinctive features of the model as to how it may differ from the usual approach, the model antecedents and the relationship to other models.

Our workshop thought that the Mcleod list ought to add a number of items including the definition of validity, or validation, because there is not common agreement on the meaning of these terms. If an assessor wants to know how well a model is validated by its developer, we ought to ask for such definitions.

Our workshop group thought it very important to add to a checklist two basic documents relating to the computer model: (1) the code itself, and (2) the code description. The modeler should provide the list of mathematical equations describing the system, and there ought to be clear pointers in the model documentation identifying which sections of the code apply to a particular section of code. This is almost essential if one is to have a reasonable chance of understanding the code.

TABLE 2

MODEL DOCUMENTATION CHECKLIST TO AID ASSESSMENT
(due to John Mcleod, Simulation Today)*

---

1. Project Information
   title, organization, contact person, project objective, project duration

2. Model Development Information
   name of model, name of modelers, purpose for which model developed, disciplines involved, data required, method of development, cost of development, availability, compatibility, extent of use

   add: choice of parameters, equation specification

3. Model Description
   classification, block diagram, program, notation, validation, reference information, distinctive features, model antecedents, relation to other models

   add: definition of validation, put pointers in model documentation to identify sections of code

4. Simulations (experiments)
   title, purpose, assumptions, experimental design, data requirements, data used, run time, cost/run, results, analysis

   add: verification history

5. Discussion
   comments, conclusions

6. Literature
   project reports, references, bibliography

---

*John Mcleod, Simulation: From Art to Science for Society, Simulation Today, No. 20, Decmber 1973.

The fourth item in Mcleod's list is simulations or experiments done with the model and he would like to see a title for each experiment, the purpose for which it was done, the assumptions used, the experimental design, exactly how the experiment was set up, data requirements, the data used, the run time, the cost of the run, results, and an analysis of the results. Our workshop thought that one ought to expand this list to include verification history; that is, how many of these kinds of experiments have been done over the complete history of the model, and how has the model been changed to reflect the learning from such experiments.

The fifth item on Mcleod's list is discussion. This item should provide for the drawing of some conclusions and insights that the model developer can give.

Sixth on the list is literature, which includes project reports, references, and bibliography information. This information is very essential for understanding the model.

The other information about the process which I argue we need for assessment is described in Table 3. The listing in Table 3 is a first attempt to itemize a set of information that modelers ought to supply about the process, if the model is a problem solving model.

The first on the list is the definition of problems. The modeler should describe what kind of interaction occurred between analysts and the decision makers. This is a very essential part of the policy process, and if it is ignored then analysts are not apt to do a very good job of selecting the appropriate tools, making the appropriate modifications to the existing models to be used, and so on. So assessors ought to know something about what kind of interaction occurred.

Secondly the modeler should provide information about who was involved, and I mean by that the different players in the process, not necessarily individuals. It would be useful to know whether interaction on problem definition happened in formal meetings, or whether it occurred by written correspondence, and lastly whether there was some refinement that took place. This set of information would give an assessor a good feel for whether the model development and use occurred as a result of a modeler's isolated attempt, as a result of interaction with the mid-management level of an organization, or if it grew out of interactions with the organization's chief executives. One should know if models were built as the result of a presidential decree that some new policy is to be adopted, the result of an organized attempt to identify key policy issues, etc. This information is very important in understanding exactly why a particular model was chosen, why certain modifications were made to it, and so on.

The second major item in Table 3 is a description of the kind of interaction which took place with the groups to be impacted by policy decisions. This information is important because it tells an assessor the extent to which the selection of models and the definition of the problem reflected the values of people who are going to be impacted by the decisions.

It may be, for example, that the analyst's first attempt to define a set of policy options leads to the conclusion that absolutely no one would support

TABLE 3

INFORMATION ABOUT THE PROCESS: AN AID TO ASSESSMENT

1. Definition of Problem(s)
   how did interaction occur, who (not necessarily by name) was involved, meetings, written correspondence, how refined

2. Interaction with Groups to be Impacted by Policy Decision
   which groups, how were definitions of good and bad outcomes determined (from each group's perspective), was research of the normative included

3. Time Constraints for Investigation
   time for problem definition refinement, investigation, interaction, testing, peer review

4. Ways Investigation (and model results) Affected Decisions

5. Description of Disciplinary Backgrounds of Analysts

a policy initiative in that direction because all principal groups believe the policy would yield a bad outcome. In this case, the process generally redefines the problem; this is a very important part of the process. And assessors ought to know the essential information about the use of models in the process--which groups were involved, how the definitions of good and bad outcomes were determined from each group's perspective, and if there was explicit research done on normative issues.

The third item in Table 3 calls for information on the time constraints for investigation. This is very crucial information. Timeliness is of the essence in the political decision making process, and much of what is done is greatly affected by the time constraint. So a modeler should provide to assessors information concerning how much time was allowed for problem definition, refinement, the investigation itself, interaction among groups, testing, and peer review group functions.

In some cases, when the importance of items listed in Table 3 are not understood and appreciated and when an organization is not structured to achieve interaction, the investigation process may take virtually all the time available and be ill-conceived because the problem was not defined very well. Assessors need to spend time examining this kind of information in order to judge the good and bad qualities of models. Results may be criticized because there are no peer reviews, and so on. So analysts need and assessors need that kind of information.

The fourth item in Table 3 is difficult but necessary. Assessors ought to know the extent to which the modeling results actually affected decisions. Assessors need some feedback about how models in particular investigations were used, and how they affected final decisions.

The last item in Table 3--and this received much discussion in our group--calls for a description of disciplinary backgrounds of the analysts involved in the modeling. Many felt that such information is irrelevant and assessors ought to be able to do the job without it. I think it adds a great deal of understanding on the part of the assessors. Disciplines are accustomed to dealing with modeling approaches in different ways and such background information provides clues to the values and orientation of the modelers.

This concludes the summary of our workshop discussion. Again, I think that the best I can say is that we got loose agreement about the essential information which should be provided in order to aid assessment. We had varied opinions about how much emphasis ought to be given to descriptions of the process as opposed to the model itself.

REPORT ON WORKSHOP 4

SENSITIVITY AND STATISTICAL ANALYSIS OF MODELS

Carl M. Harris*
David S. Hirshfeld**

1. INTRODUCTION

This panel was a relatively small one, but it was very active. The group was made up of people with diverse backgrounds, both from the methodological point of view and with respect to day-to-day professional perspectives. There were university people, Government people, and industry people, including a number from the National Energy Laboratories. As a result, we had a variety of opinions that ranged over many areas. But we were able to reach a moderate consensus on the major issues. Our discussion included possible new research areas, future ventures, and very specific individual project requirements.

It was felt that explicit treatment of the uncertainty associated with energy supply and demand projections can improve their usefulness in policy analysis. Unfortunately, relatively few energy modeling efforts to date have enjoyed this enhancement of their utility.

The workshop agreed to define "uncertainty" broadly, to include both: (1) true randomness in input data and parameters; and (2) potential errors residing in model logic, model structure, or computational procedures. Sensitivity analysis, in turn, is the formal investigation of how and to what extent changes in a model's numerical data, assumptions, logic, and mathematical structure affect (1) the results generated by the model and (2) the inferences that one can properly draw from such results.

Most energy modelers have relied almost exclusively on deterministic sensitivity approaches to uncertainty assessment (for example, see Manne, Richels, and Weyant, 1979). To quote Manne, et al., "Economic assessments typically contain benefit-cost calculations for a large number of scenarios, with each scenario representing the uncertainties as though they could be all resolved before any actual choices had to be made." For short-range models they point out that long-term randomness may have little or no impact on output measures and thus a non-random model might suffice. But even here, if alternative scenarios (varying according to some underlying probabilistic structure) lead to very different model outputs, deterministic approaches cannot provide a completely satisfactory picture. In mid- or long-term cases, some key input elements will quite likely behave in a random fashion, and thus suggest a non-deterministic strategy for assessing the range of model outputs.

It was of interest to the workshop to note the sensitivity analysis that has been done for the Midterm Oil and Gas Supply Modeling System (MOGSMS) by the Department of Energy (DOE). An assessment of the sensitivity of these models to its key input variables has been made, and shown in recent data validations and updates (for example, see Office of Energy Information and Analysis, 1976, PIES Documentation, Volume IV). Clearly, those variables specified should be the major focus for any sensitivity analysis. Of special concern might be the high sensitivity variables that may properly be viewed as stochastic (since these impact the model's basic random character). There is also, of course, the issue of model randomness caused by the estimation of the parameters from data where there is scatter about some nominal or "true" value. From a modeling perspective, this data uncertainty is as much a source of randomness as the model's stochastic variables, and its consideration may well be part of a data validation effort.

*Center for Management and Policy Research, Inc., Washington, D.C.
**Hirshfeld Associates, Potomac, Maryland

Additional sensitivity work on MOGSMS has been done at DOE by a (deterministic) analysis of changes in projected crude oil and natural gas production. Some of the key endogenous components have been selectively varied over small ranges, within each of a number of exogenous, preset scenarios.

An integral part of the National Bureau of Standards (NBS) Energy Model Validation project for DOE in 1979 was a task to determine "the sensitivity of MOGSMS results associated with the particular choices (of data values) which make up the model". This task called for a sensitivity assessment of the model by comparing alternatives to key model inputs. Particular emphasis was placed on an evaluation of model sensitivity and stability relative to such alternatives. The major finding of this task was that there indeed has been an inadequate approach to the assessment of uncertainty in the supply projections generated by the Midterm Oil and Gas Supply Modeling System. Probabilistic strategies were in fact indicated here for passing alternative scenarios through the models in a statistically correct and meaningful manner. The major technique used was <u>model sampling</u>, details of which are provided in a number of NBS project technical reports.

In a follow-on NBS study, the results of this prior sensitivity work and related tasks were applied in early 1980 to a quantitative assessment of "the quality and usefulness" of those MOGSMS results that are a part of the forecasts of midterm oil and gas supply in the 1978 DOE <u>Annual Report to Congress</u>. Specifically, this assessment is in the form of a sensitivity analysis (as defined above) of these MOGSMS results, with respect to uncertainty (or possible errors) in:

- Input data
- Logical structure
- Statistical methods

This sensitivity analysis does <u>not</u> address two other areas of uncertainty or possible error:

- Mathematical structure
- Computational procedures

Sensitivity analysis is especially pertinent for MOGSMS because its results are widely distributed, and these results appear to be sensitive to key input data elements, parameters, and assumptions. The <u>1978 ARC</u>, for example, makes frequent reference to the effects of alternative input scenarios, uncertainties in parameters and variables, and even alternative procedures for making the forecasts. This emphasis on sensitivity is not surprising, because MOGSMS is a static model, generating deterministic results that become forecasts of time-dynamic, highly uncertain phenomena.

2. PROBABILISTIC VS. DETERMINISTIC SENSITIVITY ANALYSIS

It was agreed by all that sensitivity methods are basic to effective operations research study. Presentation of the effects of perturbations of model variables and parameters is a principal way of conveying to decision makers an understanding of the environment in which they operate. Indeed, careful and comprehensive sensitivity analysis should be a backbone of any planning-oriented modeling project, because policy or decision makers usually want more than single numbers or simple sets of numbers as answers. They desire a credible assessment of the range of possible outputs and how these can vary with different actions. How, for example, can desirable end results be reached by altering some of the decision variables and what are extreme possible outcomes if circumstances do not evolve quite as originally perceived?

It may well be that a policy maker wants to set values of model parameters at any of a number of alternative levels, all of which may be feasible choices, though only one set may become the basis for action. The modeler must be able to test the effect of each such alternatives. Another policy maker may wish to choose key elements in a system from a range of practical alternatives. The effects of all such possible changes must thus be assessed in a comprehensive fashion. We refer to these kinds of uses as Deterministic Sensitivity Analysis. On the other hand, the explicit recognition of uncertainty in a model leads to a more comprehensive sensitivity testing, including that of the manner in which the output reflects the input randomness. We refer to this usage area as Probabilistic Sensitivity Analysis. In a sense, in this latter case, the same strategies for putting alternative scenarios through the model are used, but the process involves a comprehensive approach, including:

(1) determination of input probabilities and (particularly) joint probability distributions, as necessary;

(2) an experimental design for selecting specific values of input elements for analysis[1];

(3) a statistical format for analyzing and presenting the results.

A quote from Sisson (1974) reinforces these points:

> "...If the model is an optimizing form, then it is now ready for use; the data are input and the program run to produce the outputs. Most such programs not only produce the optimum values

[1] As a practical manner, an equivalent method of selection must be used in the deterministic case if model runs are costly.

of the decision variables, but also sensitivity information. This might include an indication of the costs of deviating from the optimum, and the benefits that might be derived by relaxing constraints; that is, by changing factors originally assumed to be 'uncontrollable'. If the model is only a predictive model, then one more step is needed: it is necessary to decide how it will be used to search for the best decision. The search process may be:

- try several alternatives, or
- design a simulated 'experiment' to try alternatives in a controlled way, which insures trying a wide range of them, or
- use a formal search process in which the next alternative to try is computed so as to lead to a higher utility."

It is also interesting to look back at some of the early texts on OR to see what was said about these and related questions. Churchman, Ackoff, and Arnoff (1957) considered six major phases of an operations research problem:

(1) Problem formulation;

(2) Construction of mathematical model;

(3) Deriving solution;

(4) Testing the model and its solution;

(5) Establishing controls over the solution;

(6) Implementation.

Item (5) is of special interest to us /from page 14 of the quoted citation--see also Ackoff (1956)/:

"Establishing Controls Over the Solution

A solution derived from a model remains a solution only as long as the uncontrolled variables retain their values and the relationship between the variables in the model remains constant. The solution itself goes "out of control" when the value of one or more of the uncontrolled variables and/or one or more of the relationships between variables has changed significantly. The significance of the

change depends on the amount by which the solution is made to deviate from the true optimum under the changed conditions."

3. ANALYSIS TECHNIQUES

A key technical issue is the selection of the proper methods for assessing the effects of uncertainty in energy model parameters. We are able to identify four broad methodologies in frequent use by members of this workshop.

The first and possibly most important, is called model sampling. This is a procedure for sampling from a stochastic process (even a single random variable in the simplest case) to determine through multiple trials (generally via Monte-Carlo methods) the nature and effects of a probability distribution that would be difficult or impossible to determine by standard statistical arguments. The second method, called the adjoint method, builds on procedures borrowed from matrix theory for determining the matrix characteristic polynomial and the adjoint matrix. The third method can be characterized as response surface analysis, typically carried out under a factorial experimental design. The term "response surface" refers to a formal mathematical relationship expressing the anticipated value of the dependent variable in an experiment as a function of the experiment's independent variables. The relationship is derived by statistical analysis of the results (typically) of designed factorial experiments. An effort of this type is often completed by distribution sampling of the variables found important and thought to be random. Finally, there is a fourth class of methods in which generic analytic methods are used for each calculation and presentation of rates of change of model outputs with respect to model inputs and the regions of validity of these rates. An example is standard postoptimality analysis in linear programming.

3.1 Model Sampling

We now return to the first of these four methodologies - model sampling. As noted before, model sampling utilizes the Monte-Carlo method--random sampling from a probability distribution--but does not include observation of the behavior of a process through time. Consequently, it does not quite fit the narrower definition of simulation as normally used. Model sampling, or distribution sampling as it is also called, has a long history of use by statisticians to derive empirically distributions that are difficult or impossible to derive by other means. An excellent illustration of the application of model sampling can be found in the analysis of PERT and critical path networks (see Fishman, 1973).

To illustrate this point, let us consider a PERT network. The basic idea here is to view a project as a series of activities which must be accomplished consistent with a prescribed set of

precedences, typically represented by a network-type diagram. Precedences are established by a specification of the activities which must be completed before any other may begin. Estimates are made of the time needed by each activity, and the critical-path method then locates one or more paths through the network which determine the shortest time in which the total project may be completed. PERT builds on this concept by requiring a probability distribution (generally assumed to be a beta) for each activity time. The critical-path method is then applied using the mean activity times. The stochastic character of the activities is not addressed until after the critical path is derived. The assumption is then made that the sum of the activity times on the critical path is normally distributed (by virtue of central limit theory) with mean equal to the sum of expected activity means and variance equal to the sum of the variances.

From a probabilistic point of view, two serious problems may exist with this approach, even if we can accept the fundamental critical path structure. Convergence to a central limit may not be achieved--networks are often small and made up of a dependent sequence of random variables. Second, and even more serious, is that situations exist where noncritical paths may become critical (and vice versa) because of the interaction of the various probability distributions of activities both on and off the critical path. As a result, the deterministically calculated project-completion-time distribution is biased on the low side. The degree of bias depends upon whether there are any paths that are close to the PERT-calculated critical path in length and upon the variability of activity times both on and off the critical path.

Various methods for approximately determining the amount of bias and some analytic methods for finding the exact distribution have been suggested in the literature; but simulation affords an efficient and relatively easy way of determining the distribution of the project completion time. The simulation procedure consists of the selection of one random time from each of the activity-time distributions in the network, and the calculation of the critical path and project completion time based on this sample of activity times. The process is repeated over and over and results in the generation of the project-completion-time distribution. In addition, the frequency with which different activities appear on the critical path can be determined, if desired. Although this is a relatively popular use of simulation, it is not quite the standard one.

The Monte-Carlo method, of course, is used to select times from the probability distributions of activity times; but the technique does not involve observing the operation of the network through time in quite the same sense that behavior of systems such as queueing systems are observed through time by simulation. Rather, times are selected by the Monte-Carlo method for each of the activities, and calculation of the project completion time is then an arithmetic computation. This use

of the Monte-Carlo method to determine through multiple trials a probability distribution that would be difficult or impossible to get by analytical methods is an example of what is popularly called model sampling.

Another example of model sampling is the examination of the characteristics of techniques for parameter estimation in econometric models. Let us assume we have a model comprising a large set of simultaneous, linear equations, including some terms related to each other as a sequence of variables indexed over a finite time horizon. Suppose now that some of the coefficients of this system are random variables, possibly dependent upon each other. For each realization of the set of random coefficients, there is a solution to the linear equations. Thus, the joint probability function governing the constants imputes a probability function onto the solution space. In a sense then, this is analogous to trying to solve a stochastic programming problem in which the A-matrix entries are determined as a sample from a multi-dimensional random population.

It is interesting to note that there are well-established computing techniques for handling model sampling. Since model sampling does not involve the dynamic tracking of a model through time, we do not need a timing routine in any simulation language that may be selected for such analysis. The choice of a language is instead based on ease of programming, availability of random-number generation routines, and ease of calculating statistics. In SIMSCRIPT, for example, the timing routine can be suppressed, and built-in routines for generating random variables with desired frequency distributions and routines for calculating statistical quantities are useful features from the standpoint of model sampling applications. Also, GPSS has excellent mechanisms for gathering statistics, and this can be an important positive factor in its use for model sampling.

3.2 Adjoint Method

The adjoint method is grounded in classical tools of mathematics (see Tomovic and Vukobratovic, 1972). The method can best be described in the context of control theory. Let us assume that we are dealing with a linear multivariate system formed by the differential equation

$$\dot{X}(t) = AX(t) + BU(t)$$

together with the input/output relation:

$$Y(t) = CX(t),$$

where

X	is an n-dimenstional state vector;
U	is the m-dimenstional control vector;
Y	is a q-dimensional output vector;
A	is a constant matrix of order n
B	is an nXm input matrix; and
C	is a qXn constant matrix.

If we take Laplace transforms (with argument s) of the differential equation, assuming that initial conditions are all zero, we find that:

$$Y(s) = C[\underline{I}s - \underline{A}]^{-1} B\, U(s).$$

Determination of the system transfer function:

$$P(s) = C[\underline{I}s - \underline{A}]^{-1} B$$

is however a very complex procedure.

The key issue in the sensitivity analysis context is determining the variation of the matrix P(s) in response to parameter changes. To do this, it is necessary first to determine precisely and completely the matrix of the transfer function itself. This is accomplished using relations from matrix algebra in which the system adjoint matrix plays a key role. Ultimately, a closed-form expression is obtained for the transfer function matrix, P(s), as a matrix polynomial. After P(s) is determined, it is feasible to derive relations for computing its sensitivity to variations in parameters.

This method is now being applied to a very large energy market model (LEAP), with details given to the workshop by a participant (C. Weisbin, ORNL).

4. USE OF A MODEL FOR EXAMINING POLICY ALTERNATIVES

Any single run of a model reflects a particular set of assumptions about how various factors will impact on a future energy supply or demand system. The results of each run suggest how the system will react to particular assumptions and input elements. Each model run is therefore a particular "What if?" question about the future. Some of the "What if?" questions that are asked with the model concern impacts of external forces. Other questions are likely to concern the impacts of programs that could be developed or encouraged by the Government and activities of various other players. Changes are made in the model's numerical values to reflect a particular "What if?" question and results of a run made with these changed values project the model definition of possible effects impact. Results of any individual runs defining a particular "What if?"

question have little meaning in isolation. Rather, comparison of a set of such results to one another and against some baseline results is necessary for judging the relative impact of the changes being investigated. The runs usually best suited for serving as baselines are those that assume a fairly neutral future in which very little change takes place or those which seem to be most likely. The family of model outputs to be used in comparing other results to the baseline case is one for which a reasonably large number of alternative input or parameter settings have been analyzed to cover a broad range of possible model outputs. This is particularly valuable in assessing the effect of possible surprise events or shocks on the system being modeled.

5. ON SENSITIVITY ANALYSIS IN MODEL ASSESSMENT

From the perspective of model assessment, sensitivity analysis may be thought to be in large the study of changes in model output induced by varying the inputs. As such, it becomes a primary area of concern in model assessment. Its uses may include the determination of rates of output change with respect to the input changes, importance ranking of the inputs from a sensitivity standpoint, assessment of output variability attributable to the inputs, or exploration of questions such as: What causes unexpected results, what are important inputs, and what is the output range of variability? One of the natural consequences of such statistical quantification of the variability of the model's outputs is the development of tools useful for the decision maker in employing the target model in the analysis of measures for dealing with an uncertain environment.

6. CONCLUDING REMARKS

In general, sensitivity analysis and the direct treatment of uncertainty should be integral to a model's use, beginning with sensitivity studies during model development to support the early identification of critical input data elements - those which can be characterized as uncertain and of substantial impact. Sensitivity analysis experiments support the identification and sharpening of technical assumption. As a result, they can provide guidance in the direction of data collection efforts.

At the usage level, sensitivity analysis can highlight for the end users what the critical data elements and technical assumptions are, which are often not at all obvious to people other than the model developers. Another role for sensitivity analysis is in identifying what might be called the regions of invariance in sets of model solutions. As various input elements are perturbed and different scenarios run with a given model, certain elements of forecasted future circumstances can

be found that are invariant to such changes. Formal sensitivity analysis facilitates the identification of these regions of invariance. The complementary problem, in which sensitivity analysis has a key role, identifying the regions of high uncertainty or high rates of change in solution values under imposed changes in inputs. Another area of model application that could benefit from some sensitivity analysis is the definition and selection of scenarios. Some fairly technical sensitivity analysis can provide guidance in what sets of scenarios will provide the greatest range of decision-supporting information when the various outputs are assembled and analyzed.

Finally, at the model assessment stage, sensitivity analysis has obvious roles. It is probably one of the best ways that we have available to implement systematic searches for discontinuities, anomalies, and inconsistencies in model results. It is also useful in defining the domain of applicability or reasonableness of the model.

7. REFERENCES

1. Ackoff, R.L. (1956). "The Development of Operations Research as a Science," Operations Research, Vol. 4, pp. 256-295.

2. Churchman, C.W., Ackoff, R.L., and Arnoff, E.L. (1957). Introduction to Operations Research. Wiley, New York.

3. Federal Energy Administration (1974). Project Independence Report, GPO, Washington, D.C.

4. Federal Energy Administration (1976). National Energy Outlook. GPO, Washington, D.C.

5. Fishman, G. (1973). Concepts and Methods in Discrete Event Digital Simulation. Wiley, New York.

6. Hillier, F.S. and Lieberman, G.J. (1967). Introduction to Operations Research. Holden Day, San Francisco.

7. Manne, A.S., Richels, R.G., and Weyant, J.D. (1979). "Energy Policy Modeling: A Survey," Operations Research, Vol. 27, pp. 1-36.

8. Meier, R.C., Newell, W.T., and Pazer, H.L. (1969). Simulation in Business and Economics. Prentice-Hall, Englewood Cliffs, N.J.

9. Naylor, T. (1971). Computer Simulation Experiments with Models of Economic Systems. Wiley, New York.

10. Office of Energy Information and Analysis (1976). PIES Documentation. U.S. Department of Energy, Washington, D.C.

11. Office of Energy Information and Analysis (1977). Medium-Run Oil and Gas Supply Model 1977 Data Update. U.S. Department of Energy Research Memorandum No. 78-015. Washington, D.C.

12. Pritsker, A.A.B., and Kiviat, P.J. (1969). Simulation with GASP II. Prentice-Hall, Englewood Cliffs, N.J.

13. Sisson, R.L. (1974). "Introduction to Decision Models," in A Guide to Models in Governmental Planning and Operations, U.S. Environmental Protection Agency, Washington, D.C.

14. Tomovic, R., and Vukobratovic, M. (1972). General Sensitivity Theory, Elsevier, New York (original version published in Serbo-croatian in 1970).

Enumeration of Validation Methodologies

by James Gruhl and David O. Wood
Energy Model Analysis Program
M.I.T. Energy Laboratory

1. INTRODUCTION

This report provides an enumeration of a large number of potential assessment/validation methodologies. It is a summary of some of the authors' preparations for, and the participant activities at, one half of the conference Workshop 5. The first portion of this report attempts to formalize some aspects of the assessment process. To the extent to which it succeeds, it presents a logical framework for the investigation of assessment techniques. This framework is developed, and involves

(1) defining the detailed stages of the modeling process,

(2) listing the possible assessment actions that could be made at each stage,

(3) enumerating the possible evaluations of the effects of those changes,

(4) describing the alternative models and data, and the judgments that could be a basis for evaluating those changes, and

(5) developing strategic guidelines for the selection of the most effective assessment methodologies.

Finally, using this informational framework some important assessment issues are addressed. The most significant issues identified are the areas where further research is needed. Of equal importance, but due to the enormity of the task they are left for some future review, is the discussion of the current gaps between the analytic tools needed for assessment and the available applied mathematical techniques.

2. VALIDATION HYPOTHESES

There have not yet been developed any formalisms for the study of modeling. Likewise, no scientific theories or formalisms have yet been developed that can provide guidance in the enumeration of model validation techniques/strategies. Here, in order to begin in a very informal way to develop some of the necessary groundwork, suppose that we start with a very simple hypothesis:

Hypothesis 1: Systematic validation/assessment approaches are the most effective.

Although 'systematic' (which probably should mean something like 'well- organized' rather than 'brute force') and 'effective' (including time-, cost-, and manpower-effectiveness) are somewhat vague terms, it seems likely that this hypothesis is true. Models, especially policy models, tend to be so complex and multi-disciplinary that systematic approaches would seem essential in order that important assessment issues (that may not be of particular interest to anyone on the assessment team) not be missed. In addition, a systematic approach is most likely to produce an appropriate balance of good and bad assessment issues. Finally, it seems likely that there is such a huge number of potential validation/assessment techniques that a systematic approach for sorting out those that are most effective would be very advantageous.

This first hypothesis will be left aside until the later discussions in Section 5 , however, its suggestion to attempt systemization will be followed in the pursuit of validation techniques. The second hypothesis is somewhat more subtle, attempting to identify a decomposition of every validation technique into its component parts:

Hypothesis 2: Every validation technique involves the performance of some action (perturbation) at one stage in the modeling process, a transfer of that action's effects to another stage, and then an examination or evaluation of the results with respect to some (external) basis for comparison.

It appears that the only way to evaluate this hypothesis is to examine all available validation techniques to see if they all can in fact be decomposed in this manner. Preliminary surveys show that this seems to be the case. However, whether or not this hypothesis is precisely correct, even its presumptive use will later be seen to be valuable in the process of displaying all potential validation techniques. The value of working solely with the validation components is a consequence of the likelihood that the separate enumeration of all possible action, transfer, examination, and comparative validation components will be easier and more meaningful than the enumeration of all possible validation techniques. This will especially be the case if it happens that:

Hypothesis 3: All possible combinations of validation components are in fact bona fide validation techniques.

Continuing this very informal statement of conjectures, consider one final proposition:

Hypothesis 4: There is no validation technique that could not be part of a model developer's "good modeling practice."

Of course, certification that validation techniques have actually been undertaken can only credibly be performed by some party independent of the model builders. This certification, however, is probably more

accurately classified as an assessment activity, rather than a validation technique. The consequence of hypotheses 2, 3, and 4, if they are true, is that all validation techniques may be discovered and displayed by identifying all of the potential action, transfer, examination, and comparative components that could be undertaken at each modeling stage. Such a display could be made as a matrix, with the validation components as the matrix columns and the modeling stages as the matrix rows. A first attempt at putting together this matrix is what is accomplished in the following sections.

3. MODELING STAGES

There have been any number of formulations of the various stages of model development and model utilization. Several of these formulations have been collected together in one reference (Gruhl, Wood, Goldman, 1980), with the resultant formulations shown in Figures 1 and 2. There are obviously significant variations that could be made in the development of these formulations, depending upon the type of model and the style of its development or use. In addition, there are more detailed levels of resolution that could be incorporated in these formulations. For example, the last three stages in Figure 1 have been further resolved into fourteen substages to accommodate an emphasis on verification activities that was intended in one particular research project (Gruhl, Wood, Goldman, 1980). In any event, although it is clear that any number of subdivisions of modeling stages may be possible, it is not always apparent that such subdivisions will aid in the identification of additional validation components.

4. VALIDATION COMPONENTS

As previously mentioned, this section contains the development of the matrix that has the modeling stages as rows, and for columns has the collection of some of the validation components that have appeared in the literature (the literature listed in Section 7 References and Bibliography, as well as some additional articles of less general interest). Table 1 presents the collections of the validation components concerned with the validation actions, transfers and examinations. The bases for comparison are listed separately, in Table 2, partly due to space limitations in Table 1 and partly due to the repetitiveness of their use at the various stages. It should be apparent that a great deal of judgment must be used in the selection of validation components that would be most useful for particular types of models. To aide in this judgmental selection, another area that may be worth future investigation would be the collection of those comparative techniques that have in the past been most useful for certain types of validation examinations.

Simple input/output sensitivity analysis will now be used as an example of how to go about identifying known validation techniques on this matrix. This technique resides entirely in the "Utilization" portion of Table 1. The validation action involves a systematic perturbation of "Input Scenario Specifications". The validation examination is performed on the set of values that result in the "Output Collection/Presentation." The transfer mode is more or less the default,

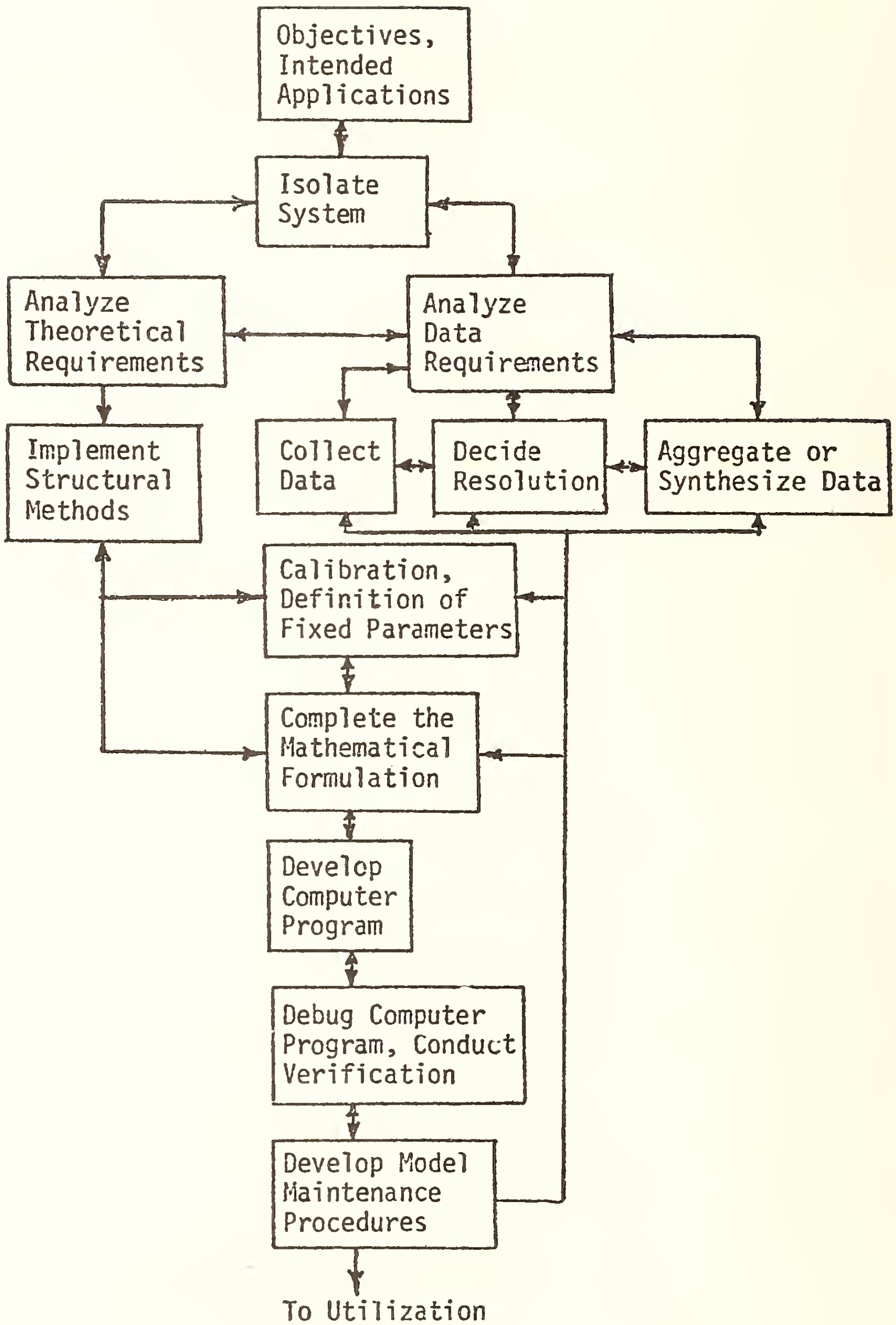


Figure 1 The modeling stages that take place during model development (Gruhl, Wood, Goldman, 1980).

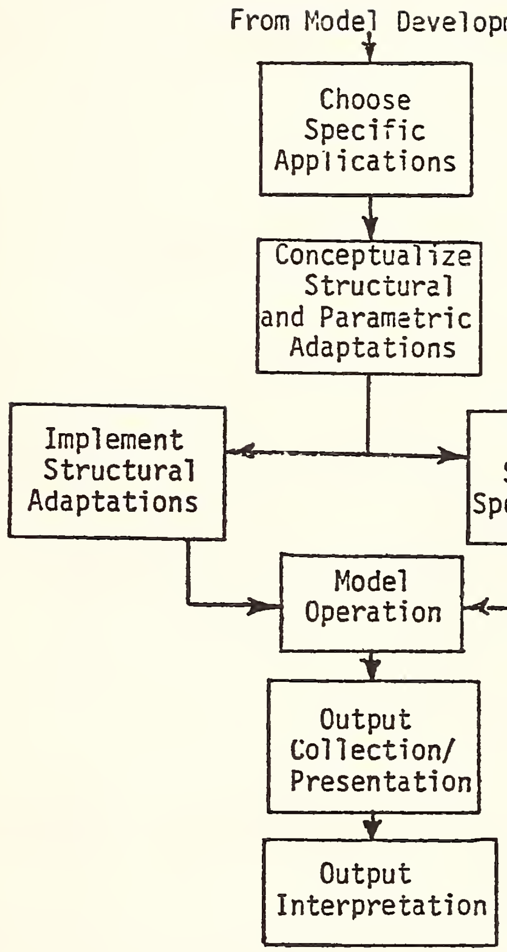


Figure 2 The modeling stages that take place during model utilization (Gruhl, Wood, Goldman, 1980).

Table 1. Validation Techniques Appropriate for the Various Modeling Stages

MODELING STAGE	ACTION	TRANSFER	EXAMINATION
	MOTIVATION		
-Objectives, Intended Applications	-surveys	-unchanged -use other model	-examine variations
	CONCEPTUALIZATION		
-Isolate System	-decompose -selective removal of variables -"free-body" diagram with one more level of detail added -additional input/output/state variables, new components	-unchanged -use other model -maturity of disciplines -examine variations	-global perspective, completeness
-Analyze Theoretical Requirements	-counter-analysis -data/structural tradeoffs, compromises	-unchanged -use other model	-evenness of detail
-Analyze Data Requirements	-survey of available data -data/structural trade-offs, compromises	-unchanged -use other model	-examine variations

Table 1 (continued)

MODELING STAGE	ACTION	TRANSFER	EXAMINATION
	IMPLEMENTATION - DATA		
-Collect Data	-systematic perturbation -selective removal of relev., irrel. or correl. data -data splitting -perturb data at point of application -error/uncertainty characterization -additional fabricat. data -new data collection	-unchanged -use other model	-correlation or irrelevance tests -examine variations
-Decide Resolution Spatial, Temp. Information	-additional detail -additional aggregation	-set new resclution -unchanged -use other model	-evenness of detail -examine variations
-Aggregate or Synthesize Data	systematic perturb. selective removal of relev., irrel. or correl. data -data splitting -perturb data at point of applic. -error/uncert. characterization -additional fabric. data -new data collection	-unchanged -use other model	-correlation or irrelevance tests -availability/unavailability of required data -examine variation
	IMPLEMENTATION - STRUCTURE		
-Implement Structural Methods	-respecification, complexify -simplify, omit -uncertainty characterization -parameterize altern., counter-modeling -fabricate or perturb structure	-unchanged -use other model	-comparison with other models -examine variation

Table 1 (continued)

MODELING STAGE	ACTION	TRANSFER	EXAMINATION
	IMPLEMENTATION - CALIBRATION		
-Definition of Fixed Parameters	-parametric perturbation, systematic or at point of application, shock analysis -error/uncert. charact.,discrete or distrib. -robust estimators diff. perference measures -constrain param., ridge regression -analytic replacement of empiric elements	-general recalib. -unchanged -recalibrate at point of appl., diff. robust techniques -use other model	-examine variation
	IMPLEMENTATION - CODIFICATION		
-Complete Mathematical Formulation	-independent reformul.	-unchanged -use other model	-examine completeness
-Develop Computer Code	-recode components, equat.	-recode -unchanged	-match code to formulation
-Verify Computer Code	-additional debugging	-unchanged	-examine completeness
-Develop Model Maintenance Procedures	-survey of techniques	-unchanged	-maturity, age of model -updating procedures
	UTILIZATION		
-Choose Specific Applications	-survey/collection applications	-unchanged -use other model	-global perspective, appropriateness
-Conceptualize Structural/ Parametric Adaptations	-survey of techniques	-unchanged	

Table 1 (continued)

MODELING STAGE	ACTION	TRANSFER	EXAMINATION
-Input Scenario Specification	-systematic perturb.,discrete distributed, or sets -error/uncert. characterization -aiming or optimizing with respect to some target set or preference -historical values	-unchanged -use other model scenario	-examine values or set of values -gradients, differences, ratios, percents -graphical displays -uncertainty examination
-Implement Structural Adaptations	-linearize, nonlinear. model simplif.,polyn. describing function -gradient search toward some opt. performance -incorporate analysis of struct. uncertainty	-simplified, complex model versions -simplified inverse of model -response surface -unchanged	-examine changes
-Model Operation	-decompose -simplify by omission, etc.	-single or mult. full model runs -different models -model inverse -operate component(s)	-examine changes
-Output Collec/ Presentation	-perturb systema. or at point of application -enforce historical outputs (to look at resulting inputs, struct. forms, param.) -target set specif. reachability, singularities	-unchanged	-stability or optimality, post-opt. analysis -direct examination of values or sets -fit, confidence or uncertainty, predictive qual. -gradients, elasticities, percentages -differences, ratios -graphical displays
-Output Interpretation	-different viewpoint -different outputs	-unchanged	-reinterpret -evaluate effects of uncertainties invalidities

Table 2. Bases For Comparison For Validation/Verification Techniques

COMPARISONS
-Comparison with other empirical models
-Comparison with theoretic or analytic models
-Against hand calculations or reprogrammed components
ACTUALITIES
-Data splitting by time, region, or type
-New data with time, experiments, or simulations; fresh historical data
-Internal and other consistencies
-Optimality or stability, discontinuity
-Realizability, feasibility checks, anomalies, irregularities
JUDGEMENTS
-Reinterpretation
-Comparison with understanding, reasonableness, accuracy
-Maturity of model/disciplines
-Examination of appropriateness, detail or perspective -- such as evenness of detail

unchanged, or business-as-usual, type of transfer. In this case it is single or multiple full model runs. The basis for comparison is the assessor's understanding about the reasonableness of the 'response' to the 'stimulus'.

There appears to be no reason why two or more actions, or in fact two or more points of examination, could not be used in a single validation technique. It is also possible, and frequently has been implemented, to begin with an action at a stage such as the Output and transfer this action backwards through inverses of the model to some earlier stage for examination. For example, outputs could be perturbed, transferred through a "model inverse" to be examined as input variations or resultant variations in pieces of collected data. This is an example of a 'non-default' type of transfer mode.

It seems as though there might be a couple of extra benefits from the matrix form of presentation in Table 1. One of these bonuses is that perhaps a new validation technique can be generated from the careful combination of different elements of this table. An example of one such newly generated validation technique, that has actually been applied in an assessment exercise, involved: (1) the use of additional fabricated data (perturbed from the point of application of the model) in "Aggregate or Synthesize Data", (2) the use of recalibration of the model as the transfer mode, and (3) the use of 'ratios of the changes' examined in the "Output Collection/Presentation." This is then a new measure of the strength of the calibration of a model.

5. GENERALITIES ABOUT VALIDATION STRATEGIES

The two most important portions of any assessment program are: (1) the quality, disciplines, and organizational structure of the assessment team, and (2) the strategy selected for the validation and other assessment activities. There are a great many different types of information that can be brought to the process of choosing a cost-effective validation strategy. The first consideration ought to be the objectives of the assessors and their sponsors. Some of these objectives (Gruhl, 1979, DOE) include:

(1) validate specific past applications,
(2) validate generic future applications,
(3) suggest model improvements,
(4) create a resource group with model expertise,
(5) establish credibility among users and observers,
(6) test model transferability, and
(7) further the art of model assessment.

Each of these objectives ought to be carefully considered, and further clarified. For example, does the model require "improvements" in its construction, utilization, credibility, accessibility, efficiency, penetrability, or flexibility.

Next there is a whole set of model characteristics that must be kept in mind:

(1) perceived importance of model, half-life of its usefulness,
(2) model size, complexity, and types of analytic methodologies,
(3) cost and time per run,
(4) operational transferability, including extent and availability of documentation,
(5) range of intended applications,
(6) previous model use and assessments,
(7) stage of model development and maturity,
(8) maturity of subcomponents, and assessment history of those components, and
(9) use of existing data bases, and the assessment history of those data.

These characteristics of the model must be carefully weighed so that the validation effort can work comfortably and effectively within the limitations of the manpower, time, and funding constraints of the assessment effort. The precise orchestration of assessments, with milestones, seminars, and so on, is a difficult but necessary task if cost-effectiveness is to be achieved.

Aside from these general assessment strategic considerations, there is the question of how to choose among the myriad of enumerated validation techniques (including individual and comparative assessments). A general strategy which has received some limited testing (in some assessments at the M.I.T. Energy Model Analysis Program) involves three steps: (1) screening to identify problem areas, (2) focusing on individual critical issues, then (3) bringing those issues into the broad perspective of the whole model and its areas of application.

The <u>broad screening</u> requires an evaluation of not only the whole model (or several of its components) but also, as an additional dimension, all (or several) of the modeling stages. The process of <u>focusing</u> involves tracking down to individual stages and components (and later equations) any counterintuitive or irregular results. The <u>broad perspective</u> is then built up through comparative techniques, with and without the resolutions or alternatives to specifically identified validation issues. This is the point at which the matrix in Table 1 offers an additional possible bonus. Assessment techniques can be identified or even generated that span large portions of the modeling process or that focus on individual stages. This can be accomplished by adjusting the span between the initiated validation <u>actions</u> and the points of <u>examination</u>.

Implementing the process of screening-focusing-perspective is still very much an art. In (Gruhl, Wood, Goldman, 1980) there is an initial effort at collecting and prioritizing techniques that have met with various levels of success in past assessments. There seems to be enough pattern to that listing to warrant further comprehensive investigation. The development and choice of the most effective validation techniques for a particular model assessment is virtually a new field and is likely to be an exciting research area for some time.

A number of research topics other than those previously mentioned were topics for discussion at the workshop meeting. Some of these future research topics included:

(1) the need for a procedure that could be used to take a set of intended model applications and from it develop the overall global perspective within which the model can be imbedded (defining the exogenous, endogenous, and untreated effects),

(2) the development of methods for expressing: data errors and the extent of synthetically developed data, structural specification assumptions and alternative structural forms, calibration errors and compromises, and output uncertainties,

(3) the most appropriate validation techniques for use on modeling components that use static optimization, dynamic optimization, simulation, or non-modeling techniques,

(4) the extent to which, and how, the assessment process can be formalized as a methodology so that assessment can to a great extent be part of 'good modeling practice,'

(5) the circumstances under which validation methodologies should best be undertaken in the context of other models, as opposed to the context of reality,

(6) the definition of the fine line between counteranalysis, which is so necessary for assessments, and countermodeling, which is inappropriate for assessments because it can create competition for the original modelers, and

(7) the creation of assessment methodologies that will give a proper balance between the strong and weak points of any particular modeling approach.

6. ACKNOWLEDGMENTS

In addition to Tom Woods, the other co-chairman of this workshop, there were several others in the workshop that made contributions to this presentation. Participants from the Department of Energy included: Cecilia Chu, Harvey Greenberg, John H. Herbert, Peter Holihan, Brenda Jorgensen, Fred Murphy, Mark Seidel, Susan Shaw, Scott Sutton, and Bruce Wing. Also from the U.S. government were Bruce Murphy and Ric Jackson of the Department of Commerce, Robert Bell from Lawrence Livermore Lab, and Robert L. Bivens of the Los Alamos Scientific Lab.

Participants from the Wharton School Analysis Center included: Susan T. Cheng, Diana Gibbons, Jill Goldman, Colleen Kirby, Lawrence S. Mayer, Suzanne Mitchell, Randi Ryterman, and Peter Tittman. Also from an academic institution was Hamid Habib-agahi of the Jet Propulsion Lab,

California Institute of Technology.

In addition there were some contributors from the private sector: Dave Andress of Science Management Corp., Richard Collier from Systems Technology Inc., Byron G. Keep of the Bonneville Power Administration, Richard Shapiro from Kappa Systems, and David Strom of Conservation Foundation. Finally, there were other occasional workshop participants who contributed criticisms and additions, all of which are appreciated.

7. REFERENCES AND BIBLIOGRAPHY

Aigner, D.J., 1972. "Note on Verification of Computer Simulation Models," Management Science Series A. Theory, 18,(11), pp. 615-619.

Apostel, L., 1960. "Formal Study of Models," Chap. 1 in The Concept and the Role of the Model in Mathematics and Natural and Social Sciences edited by H. Fredenthal, Gordon and Breach Publishers, N.Y., January.

Berman, M.B., 1972. "Notes on Validating/Verifying Computer Simulation Models," Rand Report P-4891, The Rand Corp., Santa Monica, Ca., August.

Boshier, J.F., and Schweppe, F.C., 1977. "Energy Model Validation: Systematic Sensitivity Analysis," Paper presented at the Lawrence Symposium on Systems and Decisions Sciences, October 3-4.

Boshier, J.F., Schweppe, F.C., and Gruhl, J., 1978. "Validity and Uncertainty of Predictive Energy Models," M.I.T. Energy Laboratory, Proceedings of the Sixth Power Systems Computation Conference, Darmstadt, Germany, August.

Chattergy, R., and Pooch, U.W., 1977. "Integrated Design and Verification of Simulation Programs," Computer 10, pp. 40-45.

Curnow, R., McLean, M., and Shepherd, P., 2976. "Techniques For Analysis of System Structure," SPRU Occasional Papers No. 1, Science Policy Research Unit, University of Sussex, U.K., January.

Deeter, C.R., and Hoffman, A.A.J., 1978. "A Survey and Classification of Methods of Analysis of Error Propagation in Mathematical Models," NSF, Texas Christian University, Dallas.

Deeter, C.R. and Hoffman, A.A.J., 1978. "An Annotated Bibliography of Mathematical Models Classified by Validation and Error Analysis Methods,' NSF, Texas Christian University, Dallas, April.

Dhrymes, P.J., et al., 1972. "Criteria for Evaluation of Econometric Models," Annals of Economic and Social Measurement, I, pp. 291-324, Summer.

Fishman, G.S., 1973. "On Validation of Simulation Models," Proceedings AFIPS National Computer Conference, 42, pp. 51.

Fitzsimm, J.A., 1974. "Use of Spectral Analysis to Validate Planning Models," Socio-Economic Planning Sciences, 8(3), pp. 123-128.

Garratt, M., 1974. "Statistical Validation of Simulation Models," Proceedings 1974 Summer Computer Simulation Conference, Houston, TX. pp. 915-926.

Gass, S.I., 1977. "Evaluation of Complex Models," Computers and Operations Research, 4(1), pp. 27-35, March.

Gass, S.I., 1977. "A Procedure For the Evaluation of Complex Models," paper presented at the First International Conference on Mathematical Modeling, St. Louis, Missouri, August 29 - September 1.

Gass, S.I., 1976. "Modeling and Validation in the Federal Government," Proceedings Winter Simulation Conference, pp. 558.

Gilmour, P., 1973. "A General Validation Procedure for Computer Simulation Models," Australian Computer J. 5, pp. 127-131.

Greenberger, M., Crenson, M.A., and Crissey,B.L., 1976. Models in the Policy Process, Russell Sage Foundation, N.Y.

Greig, I.D., 1979. "Validation, Statistical Testing, and the Decision to Model," Simulation, Vol. 33, No. 2, pp. 55-60, August.

Griner, G.M., Mango, J.A., and Pastrick, H.L., 1978. "Validation of Simulation Models Using Statistical Techniques," Proceedings Summer Computer Simulation Conference, Newport Beach, CA, pp. 54-59, July 24-26.

Gruhl, J., 1979. "Model Validation," Proceedings 1979 International Conference on Cybernetics and Society, IEEE Systems, Man, and Cybernetics Society, IEEE 79CH1424-1SMC, October.

Gruhl, J., 1979. "Strategies for Model Evaluation," Proceedings 1979 DOE Statistical Symposium, U.S. Department of Energy, U.S.G.P.O., October.

Gruhl, J., and Wood, D., 1978. "Independent Assessment of Complex Models," Proceedings Validation of Mathematical Models in Energy Related Research and Development, NSF, Texas Christian University, Dallas, June.

Gruhl, J., Wood, D., and Goldman, N.L., 1980. "Verification Strategies for Socio-Economic Simulation Models," Presented at the 1980 Summer Computer Simulation Conference, "Success Through Simulation," Seattle, Washington, available through MIT Energy Lab., Cambridge, MA, August 25-27.

Henize, J., 1979. "Evaluating Policy Models," Proceedings 1979 Summer Computer Simulation Conference, Toronto, Canada, pp. 550-556, July 16-18.

Hoffman, A.A.J., and Deeter, C.R., 1979. Proceedings of the Workshop on Validation of Computer-Based Mathematical Models in Energy Related Research and Development, Fort Worth, TX, June 21-23, 1978.

Hoffman, A.A.J., and Deeter, C.R., (eds.), 1978. Proceedings Validation of Mathematical Models in Energy Related Research and Development, NSF, Texas Christian University, Dallas, June.

House, P.W., 1974. "Diogenes Revisited -- The Search for a Valid Model," Simulation, 22, pp. 117-125, October.

Hsu, D.A., and Hunter, J.S., 1974. "Validation of Computer Simulation Models Using Parametric Time Series Analysis," Proc. Winter Simulation Conf.

Hunt, A.W., 1979. "Statistical Evaluation and Verification of Digital Simulation Models," Computer and Industrial Engineering, Vol. 3, No. 1, pp. 75-88.

Judge, G.G., Bock, M.E., and Yancey, T.A., 1974. "Post Data Model Evaluation," Review of Economics and Statistics, 56(2), pp. 245-253.

Kahne, S., 1976. "Model Credibility for Large-Scale Systems," IEEE Transactions on Systems, Man, and Cybernetics, 6, pp. 586-590.

Kennish, W.J., 1978. "Approach to the Validation of Computer Simulation Codes," ERDA/140901, 23p., March.

Mazur, A., 1973. "Simulation Validation," Simulation and Games, 4.

McKay, M., 1978. "Overview of Sensitivity Analysis at LASL with Coal II Example," Proceedings Validation of Mathematical Models in Energy Related Research and Development, NSF, Texas Christian University, Dallas, June.

McKay, M.D., Conover, W.J., and Beckman, R.J., 1978. "A Comparison of Three Methods for Selecting Values of INput Variables in the Analysis of Output from a Computer Code," Technometrics.

McLeod, J., 1980. "Verification, Validation and Transfer of Societal Models," Simulation, Vol. 7, February.

Mihram, G.A., 1972. "Some Practical Aspects of the Verification and Simulation Models," Operations Research Quarterly, Vol. 23, No. 1, pp. 17-29, March.

Mihram, G.A., et al., 1974. "What Makes a Simulation Credible?" Proc. 5th Annual Pittsburgh Conference.

Miller, D.R., 1974. "Model Validation Through Sensitivity Analysis," Proceedings Summer Computer Simulation Conference, Houston, TX, July.

Miller, D.R., 1974. "Sensitivity Analysis and Validation of Simulation Models," Journal of Theoretical Biology, 48(2), pp. 345-360.

Miller, F.L., 1978. "Validation of Time Series Models," Proceedings Validation of Mathematical Models in Energy Related Research and Development, NSF, Texas Christian University, Dallas, June.

M.I.T. Model Assessment Laboratory, 1979. "Independent Assessment of Energy Policy Models: Two Case Studies," M.I.T. Energy Lab, E.P.R.I. Report EPRI EA-1071, Palo Alto, CA, May.

Naylor, T.H., and Finger, J.M., 1967. "Verification of Computer Simulation Models," Management Science, 14(2), pp. 13-92 to 13-101, October.

Pugh, R.E., 1977. Evaluation of Policy Simulation Models, Information Resources Press, Washington, D.C.

Reggiani, M.G., and Marchett, F.E., 1975. "On Assessing Model Adequacy," IEEE Transactions on Systems, Man and Cybernetics, SMC5(3), pp. 322-330.

Sargent, R.G., 1978. "Validation of Discrete Event Simulation Models," Proceedings Validation of Mathematical Models in Energy Related Research and Development, NSF, Texas Christian University, Dallas, June.

Schaefer, B.M., 1975. "Model Validation Using Simulation and Sequential Decision Theory," Operations Research, 23(2), pp. 365.

Schellenberger, R.E., 1974. "Criteria for Assessing Model Validity for Managerial Purposes," Decisions Science, 5(4), pp. 644-653, October.

Schlesinger, S., Buyan, J.R., Callender, E.D., Clarkson, W.K., and Perkins, F.M., 1974. "Developing Standard Procedures for Simulation, Validation and Verification," Proceedings Summer Computer Simulation Conference, Houston, TX., July.

Smith, J.M., 1972. "Proof and Validation of Program Correctness," Computer, 15, pp. 130-131.

Snee, R.D., 1977. "Validation of Regression Models - Methods and Examples," Technometrics, 19(4), pp.415-428.

Stone, M., 1974. "Cross-Validating Choice and Assessment of Statistical Predictions," J. Royal Statistical Soc. B, 36, pp. 111-147.

Thissen, W., 1978. "Investigations Into the World 3 Model: Lessons for Understanding Complicated Models," IEEE Transactions on Systems, Man and Cybernetics, SMC 8(3), March.

U.S. General Accounting Office, 1979. "Guidelines for Model Evaluation," PAD-79-17, U.S.G.P.O., Washington D.C., January.

Van Horn, R.L., 1971. "Validation of Simulation Results," Management Science, 17(5), January.

Van Horn, R.L., 1969. "Validation," Chapter in The Design of Computer Experiments, T.H. Naylor, (Ed.), Duke University Press, Durham, N.C.

Welsch, R.E., 1974. "Validation and the Jackknife," M.I.T. Sloan School of Management, Course Notes, Cambridge, MA, September.

Wigan, M.R., 1972. "The Fitting, Calibration, and Validation of Simulation Models," Simulation, 18, pp. 188-192.

Wood, D.O., 1978. "Assessment of Energy Policy Models -- Report of an Experiment," Presented at the American Institute for Decision Sciences (AIDS) Conference, Washington, D.C., June.

Model Assessment Methodologies

Thomas J. Woods
General Accounting Office
Energy and Minerals Division
Washington, D.C. 20548

The role models can plan in policy and program discussions span a spectrum ranging from after-the-fact illustrations of decisions already made to active independent contributions to the decision process itself. Unfortunately, in the energy area the use of models has tended to be more in the illustrative end of the spectrum. This situation is due in no small part to a fundamental lack of confidence in the model results on the part of the users. The growing attention given to model assessment, of which this symposium is but one example, is a recognition on the part of the modeling community that the continued use of models as expensive illustrations of decisions is highly undesirable, both for the user and the modeler.

In our workshop session we discussed a somewhat broader perspective of model assessment. Model assessments usually focus on the mechanics of the model itself, that is, checking the data, the logic, and the methodologies used. However, there are two perspectives from which the assessments must be made.

The first is the assessment from the perspective of the model itself. This assessment is already being undertaken by many analysts. The second is from the perspective of the user, something which the workshop agreed was done very seldom, if ever. The workshop agreed that it was necessary to assess models from the perspectives of both the user and the model or modeler. The question is: how do we go about doing this?

There are three questions model assessments attempt to answer. The first is, "What does or can the model say?" The second question is, "Does the model or can the model say it well?" And the third question is, "Does the model or can the model say it 'correctly'?" In short, the user will generally ask "does it?" while the modeler will ask "can it?"

We concluded that assessing models from these two perspectives will enhance the confidence that users will have in models. And, in the end, it is the user who determines whether the modeling is something more than an academic excecise or is actually relevant to real world decisions.

These two perspectives imply differing assessment methodologies, but they are complementary to each other. Model assessment from the user perspective will tend to be rather sensitive to the particular user but rather insensitive to the individual model in the sense of methodologies, the data base,

etc. On the other hand, model assessment from the perspective of the model itself is very sensitive to the model (e.g. LP programs, econometrics, Monte Carlo) but relatively insensitive to the user. Putting together the two assessment perspectives should maximize user confidence in the model.

The assessment perspectives can be better understood within the context of overall model development and use. We broke model development and use into four general stages. They are:

--motivation, or why we do the model,

--conceptualization, or how, in very broad black box terms, we attempt to implement what we are trying to do in the model,

--actual model development, and

--model utilization.

We tried to determine which assessment perspective would dominate each stage. Motivation is assessed almost totally from a user perspective. Conceptualization is an iterative process between what the user wants and what the modeler thinks he can give him. From the standpoint of model development, model assessment is almost totally from the modeler perspective. Lastly,assessment of model utilization will take into account both user and modeler perspectives.

The user assessment techniques will be largely based on rules of thumb which provide the user an understanding of the situation the model is trying to depict. These rules of thumb tend to be mechanical in the sense that they describe what has to occur to obtain a particular result, not why or how it occurs. In a sense these rules of thumb provide the user a quasi-model of the system. With these rules of thumb, the user can assess the believability of the detailed model results. For example, any sustained production of natural gas at current levels would require the discovery of large gas fields or their equivalent at a rate unprecendented in the U.S. gas exploration history. If a model result indicated such a sustained level of production, the user would be able to discuss with the modeler/analyst what factors allowed him to be so confident about natural gas discovery rates.

This quasi-model concept raised some questions in the workshop because it appeared that we would be using a model, albeit a very primitive one, to assess another, more sophisticated model. And this is true, but model assessment techniques in use today are often comparisons of the model

mechanics with some paradigm (essentially a model) of how data should be analyzed, equations written, etc. The major problem with model assessment from user perspectives is that the "model-like" nature is more readily apparent than with assessments from the model perspective.

To date we have largely neglected to develop these rules of thumb which users, such as Congress, Government agencies, and industry, need to really utilize the results and the insights that come out of the models. To a certain extent these user assessments require the modeler to ask himself, "If I were to use the results of this model, what would I really want to know? What general rules of thumb would I use if I did not have the model to give me the detailed answers I am looking for?" In essence, the goal of these rules of thumb is to create an educated user, a user who can truly use the richness of detail and insight that models can provide to the decision-making process.

PANEL DISCUSSION ON WORKSHOP REPORTS

MR. JACKSON: I'd like to ask all of the chairmen and cochairmen to take a seat up front so we can begin the question-answer period. While they're doing that, I'd like to add that some of the slides that Jim Gruhl was using came from a very nice paper of his that appeared in the IEEE Proceedings of the International Conference on Cybernetics and Society, 1979. The report is titled "Model Validation," and does an admirable job of bring together many aspects of model validation. Are there any questions from the floor?

DR. WEISBIN: I'm Chuck Weisbin from Oak Ridge. My questions relate to Workshop No. 4 on sensitivity analysis. I did not attend that workshop. I was with another one, so my question might have already been discussed. However, I did not get clarification from the workshop summaries.

There really are two questions. The first is that you alluded to some four different types of techniques for sensitivity analysis (or an apparent battery of tools just ready to be used) and there wasn't very much discussion in terms of the limitations and appropriateness of these techniques for particular types of models.

If one talks about large-scale models--large meaning the code runs an hour using several thousand data elements--my question has to do with the appropriateness of the statistical approach. In particular, how do you screen out what elements to sample from, and so forth, and how much time does it take to do an analysis?

And I have one more question. It was also alluded that the sensitivity analysis can be used to address problems in model specification, i.e., if the equations were just out and out wrong. I know of no sensitivity analysis methodology which deals comprehensively with this case. Do you know of such a technique?

You cited input data, noise in the system, shocks, and model specification as areas of applicability for sensitivity analysis. I am wondering if you mean the latter.

DR. HARRIS: I'll take the first part of the question.

I think David Freedman did make a number of comments in his presentation addressing the issues of time, resources, and available computer fire power.

Those are all key issues. I apologize if the points did not come across, but the panel certainly focused on the relationship of these kinds of activities to available resources--personnel, project size, and model size. They are all relevant and they play a role. One cannot be terribly more definitive than that.

We did talk, for example, about screening designs. When one has a very large number of variables (forgetting everything else) and is concerned about doing some sort of fairly global sensitivity analysis, you do not really want to deal with 3,000 variables if analytic structure is really not present. If it is a question of learning which of 3,000 variables are critical, you better get to a screening design; otherwise, you are going to be working forever.

These are all issues that we did talk about. We will focus on them in more detail when we write up our report for the Proceedings. For example, in our work Dave Hirshfeld and I discuss the run time of our model compared to other models. We realized that our model was not a very big one based on criteria one normally uses. We were able to do a lot of the things we did because we had structure on our side, and we understand that that may not always be the case.

MR. HIRSHFELD: I would like to add a couple of things.

First, we had on our agenda to talk about how particular model attributes--such as sheer volume or size, as well as methodology--determine what the feasible methods of sensitivity analysis were, if indeed any existed. In other words, we wanted to get into those issues of how you can do it in individual settings, and we just flat ran out of time to really delve into it. That was why I didn't address it.

We were certainly aware of the fact that in a given setting--like an enormous LP model or like LEAP, to take an example of a large-scale model of another kind--that considerations such as volume make the doing of this stuff a lot harder than just a good presentation would suggest.

On the other hand, I guess we found out that with computer hardware and software technology advancing, as well as methodology advancing, that in fact you can tackle, in some settings, bigger mathematical constructs than you might think initially.

DR. HARRIS: Let me further respond to the question about model specification. We really have no technology available to look at the role of model misspecification. Clearly, that is a source of uncertainty. We recognize that, and I think this must ultimately be included with data uncertainty and the others.

There are, of course, uncertainties that arise from the fundamental randomness existing in model variables and parameters, and certainly the model specified plays a role in determining how much there is. I have absolutely no idea how to measure such effects, nor do I really want to push too hard on that, except to agree that one has to recognize it. I think it is indeed important to remember that model specification or misspecification does play a role in uncertainty, but the quantification of such a concept would be difficult at best.

DR. LADY: I have questions for John Maybee and Lambert Joel.

Did I understand that model confidence, as it evolved in your session, was an ordinal idea? Did you say that? If you didn't say that, I don't have a question.

MR. JOEL: No, no. As a matter of fact, that question arose and was essentially ignored.

DR. LADY: What did John Maybee say? Didn't John just say that that was so? Or not?

DR. MAYBEE: I said that this question had been brought up by Alan Goldman, and it had been discussed very briefly. I was disappointed that we didn't have further discussion of it, but somehow or another the discussion drifted away and we never got back to it. I myself feel that all there is is an ordinal concept, but--and I had the impression that Alan felt the same way.

DR. LADY: All right, now you've said it.

DR. MAYBEE: If I'm misquoting Alan, he's here to defend himself.

DR. LADY: It would seem to me if you took that position that whatever you mean by confidence could not be related to specific results, it would have to be only related to a comparison of results. So, for example, if I had a large annual report and an ordinal notion of confidence, I couldn't tell anybody about the confidence, whatever that means in terms of those results, without having some strawman or something to compare it with. Wouldn't that be right? It seems like a fearsome limitation.

MR. JOEL: There is a fearsome limitation, but I think you misstated it, in a sense. We don't really believe that it's possible to specify some quantitative measure of confidence, or confidence-worthiness of a model.

It's necessary, of course, to have two models to compare, if you want to rank them ordinally, but you don't have to look at outputs at all. You could certainly rank two models--or two anything--on the basis of what we think of them as a result of any kind of test you want to impose on them.

If I want to rank--if I think that red is better than blue and you give me two objects--I don't care if they're models or whatever--I can say that the red one is better than the blue one, and that's an ordinal relationship and it doesn't require any outputs.

DR. LADY: Well, let me put the question another way. Was there anything that happened at any point during the workshop that would be promising in terms of some notion of confidence actually being brought into being in a way that would do anybody some good, e.g., someone who actually had a report and wonders about its quality?

DR. MAYBEE: Yes, there was, but I forget all the details of it, but John had a,....

MR. JOEL: John Dearien.

DR. MAYBEE: John Dearien had an idea which he had explored with engineers in his company, and had attempted to develop a scale with the help of an industrial psychologist. I got the impression that he hadn't succeeded in convincing anybody that this could be used.

MR. JACKSON: John, you have a question or a comment? Would you identify yourself first.

MR. DEARIEN: John Dearien, EG&G, Idaho.

The research that John Maybee was talking about was involved with the assessment of thermal hydraulic computer codes. What I did there was to take some generic-type comparisons of data and computer code predictions and had 150 engineers, who were supposedly competent in assessing the comparisons between predictions and data, and had them rank it. I then tried to compare this with comparisons that we were making of codes and data.

The results I got were taken to an industrial psychologist or behavioral psychologist. He told me that basically this was typical human behavior and it was of an ordinal type--the only way you can express it is as an ordinal-type comparison.

I was able to take these results, based on his recommendations, and plot them. I got a very good comparison between people's concept of worth and the actual numerical data that I got.

As John said, I was not able to convince my regulatory agency that this was feasible, but the comparison was very good. Basically, what it said was that if you get this type of results and get enough back-up data from competent people, that you can expect a certain percentage of people to agree that this is an adequate representation of the data. The next step is getting the policymaker to go along that this means something, but it can be done.

MR. JOEL: Let me respond to the question you raise, particularly in light of what John just said.

Although we were very unhappy with the notion that confidence was a public relations kind of problem, I think it is important to recognize that it exists. If you believe that the results of analysis are useful and should be used, and you are faced with certain kinds of prejudice and you have concluded--as I think John and I have--that confidence in a model has a lot less to do with the objective virtues of the model itself as they have to do with people's impressions of them, then although you might find it distasteful, you need to sell the model or the analysis, in the sense that you are putting something over on a client. I made a list of remarks that were made in our session yesterday, which I looked at this morning, and realized that they were giving me a message of some sort. I'll read the list very fast. This is a list of things that affects confidence in a model. They're really unjust, in a sense.

Number one, confidence in a model can be affected in a positive way by a bandwagon effect. If a group of users had to use the model because they have nothing better available and this now becomes part of the lore of the model, subsequent users will be found just because the model existed and was used and therefore must be good; or at least gives the impression of being good.

On the other hand, there's a whole collection of things that can deleteriously affect model confidence. They really don't have a hell of a lot to do with the objective virtues in the model. One of them looks as if it might be objective, that is that there is a single assumption in a model development which is palpably false, or can be demonstrated to be false. If it's 1 out of 50, and the rest of the model really is very good, people will just pay no attention to the model once they've learned that it has this tragic defect.

Another aspect is what I called tendentious tinkering with a model by a user other than the developer, or other than the sponsor's people. We have an anecdote: a model was transported somewhere else and that the second user didn't really understand how to get it running, but after a lot of work, did. He additionally made, sub rosa, some changes--surreptiously he made changes in the model structure, knowingly. Then, not admitting it, got different results, different outputs, and different conclusions than the original user, and published these results and said, "Well, we used your model and we didn't get the same results." That sort of thing, once it's aired, will also reduce confidence in a model.

I won't waste a lot of time going through this. What I'm suggesting is that there is a need for good assessment of models. There is also a requirement to pay attention to the non-objective factors that influence the degree to which a model is used and respected; if you have to call it public relations, that's too bad, but you really have to pay attention to that sort of thing.

DR. LADY: You guys are sure that you want to have all of this relate to the states of mind of individuals, rather than the nature of the mind.

MR. JOEL: Yes.

DR. MAYBEE: I'm not sure I'll buy that. I think Lambert is more willing to think along those lines than I am.

DR. LADY: Well, let me ask a corollary question of everyone. In any of the workshops, did anything come up which struck you as being appropriate or indicated that we needed something that would be appropriate for helping somebody understand what they've got, quality-wise, when they get model results?

MR. HOLLOWAY: We didn't discuss this in our workshop, but maybe the following would help clarify it.

I've seen some work done to expose the explanatory power of the model. The procedure is to classify a set of exogenous variables, the things that are specified outside the model, as opposed to those that are endogenous, and then do some perturbations on the model, and analyze the results.

One takes the key outputs of the model, perturbs the inputs--the exogenous variables as opposed to the endogenous ones--and you get a measure, some aggregate measure of the explanatory power of the model.

I don't know if any of the workshop groups discussed this kind of approach. Results of such experimentation is very useful information to have in assessment.

MR. LADY: That's sort of like writing scenarios, in a sense.

MR. HOLLOWAY: Yes, in a way, but you tend to get some standardized measure like percentage changes in the variables that are exogenous to the model and then look at the results and see if that's a lot more important than the things that are supposed to be explained by the model.

I'd like to know whether anybody discussed that kind of a thing, or not.

MR. WOODS: We didn't discuss that, but we said that you had to assess the model from two standpoints: from the standpoint of the user, to give him confidence in what the results are, and from the standpoint of the modelers, to give them the confidence to use the model subsequently.

At GAO, we continually get bombarded by Congress with questions like, "Somebody just did this analysis. Is it reasonable? Somebody else says that there's a range of scenarios in this analysis. What is the correct or reasonable answer?" As a result, we are developing a codified set of rules of thumb of various issues like production of oil and gas, residential energy demands, and so on.

They are not meant to give precise forecasts. They're meant to say, "Here's how the system fits together in a mechanical sense." In other words, A plus B plus C. No causal relationships. If you put these things together the way they are, this is what's happening. If somebody comes up and presents an analysis to Congress, then the congressmen have the ability to use these rules of thumb to begin to ask intelligent questions of the analyst to determine why he came up with a particular conclusion?

For example, what if somebody says, "We can hold oil production level through 1990 by increasing production in the lower 48?" There are certain rules of thumb based on the physical realities of how oil is produced which, if you educated the user to look at, provides him the ability to identify very quickly how such a production level could be achieved. He would be able to say, "I see you're going to rely very heavily on enhanced oil recovery, or your're relying very heavily on the Overthrust Belt. Why is it that you are so confident about these factors?" As a result, on that basis, he has the ability to begin an intelligent dialogue with the analyst.

I think the reason that there's a question of confidence in models goes back to what Lambert Joel says. He doesn't like the idea of selling the product. Unfortunately, none of the models that we have here that we're "peddling" are being peddled as elegant artistic forms for their own sake. Art for its own sake doesn't justify multimillion dollar contracts.

These people are spending good money to hire us to use models because they expect something useful, and on that basis we have to assess the models on the basis of making the models useful for them. This is not necessarily confidence but we're trying to give insight, not just numbers, to give them the ability to truly understand the insights that are coming from those results.

DR. LADY: But there isn't a "them," necessarily.

MR. WOODS: There isn't a what?

DR. LADY: A "them." In other words, the annual report, which is the major product of the EIA, must consume easily more than half on a annual budget of all the resources, and the audience is indeterminate.

The problem is to have something which communicates generically what these results represent, rather than get into a tailored discussion of a particular individual and to prompt his state of mind on the confidence data.

This is not a staff function. It's an information function.

DR. GREENBERG: Harvey Greenberg of DOE.

I was in John and Lambert's session, and I have some recollections of the discussion that addressed George's question. We essentially brought up the question that George raised on Monday and at least two other occasions over the past year and a half about the nameless, faceless clientele that aren't really using the model, but are using tables of numbers that were published that came out of the model.

In various forms George has asked, "Is there any possibility that every number could be accompanied by some indicant to the confidence interval? Is there anything like that that's possible?" This would be the final confidence--quantitative confidence--measure, if you could do it, that would address the tables of numbers.

Well, in this session, as I recall, we brought that up. First, we need to consider the clientele who are nameless, faceless people and who are basically using tables of numbers, rather than the model directly. They aren't looking for insights from the model directly, but are looking at the numbers and maybe will read the texts surrounding those numbers--maybe. I think we reached a conclusion--mostly by silence--that suggests that we don't know how to do quantitative confidence; so the answer is no for this client group.

Then there's another clientele that are identifiable. These are people--so-called users, clients, decision makers, policymakers, whatever they're called--for whom you're going to use the model directly. You're going to try to provide an insight. The way you get to use the model depends on how a question is asked. That is, rather than "What is the price of oil in 1995 and give me some measure of confidence around that number?", the question should be "Which of the Alaskan pipeline proposals seems like the best thing for us to be engaged in, and when do we negotiate with Canada? What should the timing look like? Is 1985 a good year to bring it in? Or is 1990 better?"

So it is in the form of the questions. There the confidence issue has to do with how well the analyst has used the model. There's a lot of talk about whether you can analyze the model apart from the analyst. I don't think there was a consensus reached, and we did dwell on the debate that might have ensued. The Chair kept us on the track.

Well, we wound up producing a checklist which, if it were followed, third-party assessors could look at that checklist and see what you did in operating the model and in establishing controls over the model. The data base administration, the algorithmic integrity, and all the other things that were listed and summarized this morning were actually discussed at great length yesterday.

I think the summary stops short of some of the insights that were gotten out of the discussions, because almost be definition summaries do that. But the expanded version that I hope will get into the proceedings will convey how we think we have advanced our knowledge of confidence building for that second clientele. For example, if annual forecasts are accompanied by stacks of documents and appendices that describe the operational controls and the other measures taken to support the generation of those tables, then the presence of those documents will be confidence building for the nameless persons as well as the identifiable clients.

But the answer to the quantitative question, as I recall the discussions, is that we can't do that. We can't say if the table number is five, we could have something like 95 percent confidence that it's really between four and six. I believe our consensus was we can't do that, but that we can build confidence by these other things.

DR. LADY: Well, we ought to be able to say it's five and not 25,000. There has to be some cutoff. I don't accept the idea it's impossible.

MR. OSANZ: George Osanz, Science Management Corporation. I, too, was a member of this workshop, and I believe that although we could not identify a quantitative measure of confidence, in many cases we could measure a quantitative value for disconfidence. That is, we could exclude quantitatively a lot of bad models, but anything that passed through the exclusion process we would be at a loss to assign a quantitative measure.

DR. MAYBEE: In what you said, George, I detected some misinterpretation of what Harvey said. I don't think Harvey is trying to imply that it was impossible to obtain confidence intervals, we just don't have the technology to do it at this point.

DR. LADY: I interpreted what Harvey said to mean that we wouldn't.

PARTICIPANT: Harvey said they can't do it. Harvey did not say it wasn't possible to be done--although I think it is, personally, but I'm not suggesting the group said that.

DR. GREENBERG: What I'm suggesting is that as far as we know, except for ranges that are useless--certainly if it's over five, I think we can say that it's not 25,000--and if that inspires confidence in our forecasting, then I question the intelligence of the reader.

I don't think we interpreted your question as addressing the useless range category. I interpreted the question to address whether there are useful ranges of confidence that could be quantified, akin to a confidence interval. It's not between minus infinity and plus infinity, or whatever the machine precision range is, but rather some useful range--between four and six, maybe. Something like that.

But what I did say a few minutes ago is that we don't know how to do it to get a useful range with a number that is defensible. That it is beyond our collected state of the art to do that.

MR. JOEL: I'd like to add to that. Just so you won't be throwing it at me, George.

To say that we can't find a quantitative measure of how good a number is doesn't mean that we propose to--

DR. LADY: Harvey says we can, but only useless ones.

MR. JOEL: That's true. I believe that's true.

On the other hand, that doesn't mean we're going to leave a space blank for the number in a report. If you're dealing with a physical constant, you frequently are required to write down how you measured it. Numbers that appear in the report should be accompanied by whatever information you can give to measure.

Then the reader can make up his own mind how to interpret that as a quantitative measure of how good it is.

MR. GRUHL: Let me just give one example about a case where in fact some confidence limits have been developed.

There's a technology assessment model that's been used by EPA and some utilities and oil industries which essentially involves mass balance as its functional forms, and therefore, only has uncertainty in the input data and in the coefficients of the model. And these have been sent through as distributions, so that you actually get a distribution on the output. That particular model has been useful for a use in reverse, where you want to do some R&D funding exercises and want to find out where the critical uncertainties are and what the coefficients of the inputs are that are causing the uncertainty.

But I think where the methodological problems occur is that we don't have any reasonable techniques for characterizing uncertainty in structure. I think almost all of the policy models have enormous uncertainties in the structure and I think that's where the methodological block is at this point.

DR. GREENBERG: Let me suggest a way of doing it and then tell you why it's not really a way of doing it. It's only a way of appearing to be able to do it.

We could randomly sample the universe of inputs to our model--like MEMM or PIES, whatever you want to call it--and we could run a Monte Carlo and sample the output. We can even write computer programs to process the output and produce a report and thereby obtain things that would appear to be ranges of uncertainty around output numbers.

The reason why that's only apparently good, and not really good, is that you can't really do that because the number of numbers and the correlation among the number of input numbers are a big factor in designing that experiment, and you really can't do it right.

If you understood all of the correlations really well for the thousands of numbers that go in, I suggest you try them in the model in the first place. If you want to go public and say this is a 95 percent confidence interval, the knowledgeable reader presumes that's a technical phrase and has a very specific meaning and it's not intended to be colloquial. Now, if you want to say well, it's kind of, sort of in this range, then you have to define "sort of."

A lot of the reports that we're talking about, I think, are picked up by a technical readership and may be used in work that they do, just like we do census data. Other people may use EIA's reports in that way. Because it's a technical readership, you can't have some ambiguous notion of probability or confidence interval; it has to conform to understood concepts. I don't think you could produce that, except in a misleading way.

DR. HARRIS: I agree, strongly.

MR. WOODS: I would add one thing. If you really wanted to try to inspire confidence, when you publish your results, just bet your job on it. This would demonstrate the extent to which our confidence is there. I think we are mixing up the confidence required for energy models with the confidence that a user has in a model used to design this building.

There's a fundamental difference between the models we're talking about, namely energy models and the models that designed this building in terms of the kind of world they deal with.

If, for instance, I want to model how I'm going to send a ship to the moon, I build a model to describe how it will get there, and I know that everything is fixed or determined. If my moon environment were like the energy environment requiring an equivalent moon-energy or policy model, I might get my policy rocket to the moon only to find that somebody picked up the moon and moved it 50,000 miles. As a result, my rocket would miss the moon.

There's a fundamental theorem in thermodynamics that says entropy tends to increase. Unfortunately, in a human system, you can get one person who reduces entropy, and it completely turns things around. And that's the one thing that you can't give a confidence interval to, and that's why I think I agree with Harvey's problem that you can't really develop confidence intervals because an individual or set of individuals could decrease entropy in your system.

MR. HIRSHFELD: In that regard, if what you're trying to do is inspire some measure of confidence in a reader, if you can assert comparable confidence in the assumptions not changing, or in a real world failing to conform to your assumptions--and I don't think we can ever do that--then all you can say is if these assumptions don't change--and I can make no statement about the likelihood of those assumptions not changing--then I can impute this kind of confidence.

Having said that, if you could ever get to that, what have you said? The whole point is that any forecast can be knocked down if the assumed conditions are proven not to happen in the real world. Any forecast.

DR. LADY: Well conceivably that means that we have to be simpleminded in terms of what we do, rather than giving up entirely. There are some things we can explain and some things we can't.

I would be interested if we could explain the ones we can. The idea that we can't explain any I find to be disappointing.

DR. RUBIN: I just have a comment about this discussion.

We talked around these kinds of issues in our workshop as well, and we also didn't come up with any conclusions. It occurs to me, as I listen to this discussion, that there's a difference between doing analysis and communicating analysis to a specific individual where you can work with him on a one-to-one basis and doing something like the annual report, which is for a more faceless mass of users.

The question that occurs to me--and I certainly don't know the answer and don't have a position--is that on the one hand, where you communicate directly with a user and provide him with insight, then this analysis for a policymaking purpose is a useful activity.

But on the other hand, I am not so sure that the annual report type is, and I raise that question. I don't know if it's an appropriate activity for all these resources to produce an analysis which is unfocused.

DR. WEISBIN: I guess I want to go counter to the bandwagon and reverse direction a little bit. First, let me say I think I agree totally with Mr. Gruhl (and I hadn't previously met him). However, after he started talking we had five speakers in a row who attempted to refute his original position. Hence, I'd like to come back and say what I think was said, since I think several things were misstated in the follow-up replies.

First, Mr. Gruhl said that he thought that we could estimate a component of the uncertainty--that which is due to data--and propagate it through codes. Where the technology is now weak is in the uncertainty due to the model specification. I agree with this position totally and will try and review why I think that was argued with.

First it was said that the codes are very large and there is no way that we can get sensitivity to all of the data involved. I assert that I believe that is incorrect. There are ways (e.g., adjoint perturbation theory) to get sensitivities to the entire data field. We can discuss that later in more detail, if you desire, at another time.

It was then said that there is no way that you can get correlations in the uncertainties in the data. I also disagree with that statement. In fact, if you can estimate uncertainties in the data, I would ask what is your confidence in the nominal number to begin with? It is a formidable job to get those uncertainties and correlations, make no mistake about it, but there was an implication of technological infeasibility, which I think is absolutely incorrect.

Furthermore, there is this whole bandwagon movement that to be quantitative is taboo. What do we have left? Let me take the opposing position and say that if you cannot do this, what you have left may not be worth very much.

In this conference I've heard just lots and lots of words about collecting data and analyzing data and models, and it's very, very hard to walk out of the meeting with something that you can call tangible. For example, when would--if there were no monetary constraints--anyone on the panel vote when an evaluation could be declared finished? Would anyone know, once standards were met? It seems that our experience is that the only time that you can do something which is constructive and finite is when you can begin to estimate uncertainties, if you will, quantitatively.

Do not confuse that with the fact that uncertainties do not have to be small. If your estimate says "I don't know very much about that data," these could lead to very large uncertainties. If you just think about correlations, you can begin to write down what is known about them. The values of the numbers have nothing to do with the fact of whether you can get a hold of these numbers and whether you need them.

So I'm just saying in summary, I do not want to take the total opposite position, but I do believe you can estimate uncertainties due to data, and I do believe that the main weakness in the field now is still in the model specification.

MR. JACKSON: Thank you. There are a number of questions here--we could go on--but we've run out of time. I'd like to thank all of the panel members for appearing before us this morning, for taking the lead in chairing the discussions last night, and for going through the pain and agony of preparing the presentations late last night.

DR. GRUHL: Can I just make one comment.

Not to bicker about semantics, but I think the argument about correlation of uncertainties can be viewed as a structural problem within the model, as opposed to an input specification problem.

MR. JACKSON: Now I'd like to close the morning's session. Thank you very much.

REFLECTIONS ON MODELING AND MODEL ASSESSMENT

by

A. J. Goldman (*)

1. Introduction

It was a little over three years ago, in April of 1977, that I had the privilege of welcoming to the National Bureau of Standards the attendees at our workshop on the utility and use of large-scale mathematical models [26]. That was a profound pleasure, because of my strong and continuing conviction that much greater attention needs to be devoted -- both by the professional modeling community and by the user community -- to working towards better understanding of and better practices in the development, assessment and use of large-scale decision-aiding mathematical models.

Eighteen months later, thanks to the support and initiative of our colleagues in DOE's Energy Information Administration, I found myself with the equally pleasant charge of greeting the participants in a second workshop, one devoted more specifically to issues of model validation and assessment [28].

It is good to be here today on this third occasion, though now in a valedictory position on the program rather than my previous salutatory post. One temptingly simple approach, to estimating the rate of progress in the field we're discussing, would be to compare the title of the previous workshop with that of the symposium we're now concluding. The only change is that the phrase "Validation and Assessment Issues" has been replaced by "Validation and Assessment." The "Issues" have vanished! Can they all have been resolved since the January 1979 date of the last conference?

Not likely. The area of model assessment and analysis is still very richly endowed with deep and difficult and challenging problems, perplexing issues of all kinds -- practical, theoretical, conceptual, philosophical, even ethical. And I hope to touch briefly on a few of them today.

(*) Department of Mathematical Sciences, The Johns Hopkins University; Center for Applied Mathematics, National Bureau of Standards. The text contains the author's personal views of the topics discussed, rather than any kind of "official position."

But fortunately, settling of these foundational difficulties is not a prerequisite for beginning and moving along smartly on a number of useful and necessary tasks. I think most of us sense that there has been a real start towards meeting the call formulated by Martin Greenberger and his fellow-authors in their 1976 book Models in the Policy Process [34; p. 399], a call for "the development of a new breed of researcher/ pragmatist -- the model analyzer -- a highly skilled professional and astute practitioner of the art and science of third-party model analysis." We can see the slow emergence of a concept of model analysis as a serious professional specialization, rather than an activity carried out as a temporary aberration or diversion or perversion or short-term special assignment by persons whose true calling is model development. I suspect, though, that for a long time to come we will have a pecking-order in the modeling community with status-rules including "Those who can, model. Those who can't, assess," and "It is more blessed to model than to evaluate." An interesting question, which might well be addressed at a subsequent conference in this series, is that of what education is appropriate for the new breed of analyst; this seems even more difficult than the corresponding question for model-builders (cf. [63]). Training like that offered in [42] might well be a requirement in such a curriculum.

2. Documentation and Standards

One of the areas in which it was especially easy to perceive widespread failings in current practice was that of model documentation. It was a sobering reminder of the passage of time to find, on checking the files, that my own first public diatribe on this scandalous subject took place more than fourteen years ago [32]. But of course that was not the earliest such jeremiad; here, for example, is a curiously current-sounding quotation from a War Gaming Symposium held in 1961 [56]: "There are already some good simulation models which are essentially lost because the teams which created them were dissolved before making sufficiently detailed records -- waste also occurs when proper records are not made because new teams must resolve many of the same problems over and over."

Quite naturally, there is talk of documentation standards, and a body of work underway (e.g. [25, 27]) which may lead to their development and use. You will recall, no doubt, the passage near the end of George Lady's address on Monday in which George bared his fangs and observed -- in charmingly low-key fashion -- that standards and ideals of good practice only really take force once some profferred projects or reports are actually rejected for failing to meet them. The modeling

community does recognize a need for greater discipline about documentation [26, pp. 211-212; 29], though perhaps preferring greater opportunity and incentive for self-discipline. But there is also an entirely reasonable concern about imposing a set of inflexible and rigidly-administered standards upon any part of a creative and varied process like modeling (e.g., Greenberg [33], Nissen [55; p. 282] on a "danger of totalitarianism"), and in particular some nervousness about the active role of the National Bureau of Standards in this model-assessment field (cf. Hogan [37]).

It is therefore worth saying, explicitly, that the National Bureau of Standards is not populated by wild-eyed "standards freaks." At the NBS the scientific staff is terribly aware of how demanding it is to develop standards well, with a solid technical rationale, and in a cooperative way that truly merits and receives the acceptance and appreciation of most users. The job becomes even harder, in many ways, when what is sought is a functional standard (e.g., "noise above a certain threshold shall not pass through the apartment's walls") rather than a material or structural one (e.g., "the walls shall be constructed of Material X and shall be at least yea thick"). But functional standards are what's really needed, if ingenuity and creativeness are not to be restricted.

With few and partial exceptions (e.g., Shaw [60]), the modeling profession has not as yet seriously approached the issue of model documentation standards from this rigorous "functional" viewpoint. Doing so requires more than reaching a consensus that all good modeling efforts, of a certain size and complexity, should produce such-and-such documents at certain stages of their progress. It would involve, first, thinking through very carefully what it is that documentation is supposed to accomplish in terms of transfer of information, ability for new people to get a model up and running, ability to modify it, ability to understand its assumptions, etc. Then one wants to develop means (measures, and measurement methods) for determining to what extent a given body of documentation succeeds in performing these functions. One needs a deliberate articulation of alternative documentation and training approaches, and testing of these approaches against the performance measures so that design guidelines can be built up.

I do not mean to imply that current efforts, to apply good sense and professional judgement to the development of interim documentation standards, are in any way inappropriate. But I do suggest that the type of conceptual, theoretical and experimental research program described above should lead to a next generation of better-based and more flexible standards, and also to some fundamentally important knowledge about communication of and through models.

Let me give just one example (perhaps not a terribly good one) of the kind of problem I see lurking in the documentation thicket. One injunction on everybody's list of documentation advice is: list all significant assumptions of the model. And so let us suppose that our model assumes a certain relationship and has the nice simple form $Y = AX$. Now, mathematics is an extremely powerful language for expressing assumptions. Once Y and X have been defined and explained, when you write down that equation it states exactly what is being assumed, nothing more and nothing less. But it seems to me that this leaves out a kind of information that some users would need, namely what might you plausibly have assumed <u>instead</u> (but didn't)? Was the real point that you assumed a zero intercept? That the relationship was linear rather than quadratic? Rather than exponential? It appears by no means trivial, in a document of finite length, to convey to varied readers with varied needs the <u>significance</u> (rather than just the statement and justification) of the model's containing one particular set of assumptions rather than some other. (An interesting early example of assumption-exposition is given in Chapter 5 of [35].)

A change of subject: during the last few months I've had occasion to do a little reading about university libraries. Standards for such libraries are set by appropriate professional associations, and these standards display an interesting gradation [66] worth reporting here.

First, there are <u>accreditation</u> standards. They are regarded as minimum thresholds: a library not meeting them has no real right to call itself a college library at all. Next, there are <u>benchmark</u> and <u>diagnostic</u> standards. These are supposed to represent roughly the canons of good practice -- what is generally achieved by decently-regarded university libraries.

The third and most intriguing category is that of <u>projective</u> standards. These represent, and in fact go somewhat beyond, what is done by the "forefront" academic libraries. They are goals, desires. They're intended to stretch existing practice, to stand ahead of it far enough to be inspiring but not so far as to be hopelessly discouraging.

Perhaps such a hierarchy of classes of standards, each class with its own distinctive role, is worth keeping in mind as we think about the development of standards and guidelines for the planning, conduct and documentation of modeling and model-using activities.

3. Targets, Unicorns, Elephants and Icebergs

In the process of learning about model analysis and assessment through real experience rather than anticipatory speculation, we find ourselves in confrontation with some sticky problems -- problems we were aware of earlier at an intellectual level, but sensed to be so unpleasant that we tended to dismiss them from attention until actual practice plunged our noses smack into them. I'm going to mention just three of these now, and I'll refer to them as

-- the Moving Target Problem
-- the Unicorn Problem,
-- the Elephant/Iceberg Problem.

3.1 The Moving Target Problem

The moving target problem has been mentioned, sometimes under that very name, in several past papers in this workshop-series (e.g. Lady [47, p. 9], Wood [68, p. 40], Richels [57, p. 176], Baughman [5, pp. 200-201]). The issue is that of what procedures are fair, practical and appropriate in trying to evaluate a model that is so impolite as to refuse to stand still in its evolution so the assessor can get a firm grip on it. Perhaps one might call it the "Protean model quandary"; Freedman [23] speaks of "model fluidity".

Well, that sort of problem has a long history among hunters and military marksmen, and there is a well-developed procedure for dealing with it. That procedure will surely come to mind if you recall how modern computers arose from the needs of anti-aircraft gunnery in World War II.

The way you handle a moving target is to lead it, and that principle is the basis for my proposed solution of the problem. In traditional fashion, I lay it out as a 3-step process:

(a) Gather and analyze data on the evolution of Model M prior to the present time t_0. Also gather and analyze relevant data on the evolution over time of past models of similar type, past modeling efforts by the same group, etc.

(b) From this rich data base, develop a model M' predicting the evolution of Model M between now, $M(t_0)$, and the time (t_a) when the assessment is due.

(c) Apply your assessment techniques to the predicted version of $M(t_a)$.

For ease of presentation, I have omitted some minor technical complications from this description. One of them is the need for a proper assessment of step (b)'s Model M', a thought suggestive of the whole fascinating field of meta-assessment, complete with Russell-type paradoxes about the class of all assessment models which do not assess themselves, etc. (cf. Smullyan [62]). But you get the general idea.

Homework problem -- essay type. Is there something serious and worthwhile in this facetious prescription? (I really believe that there is.) Please turn in your answers before the next symposium.

3.2 The Unicorn Problem

To introduce this problem, let me tell a little story. Imagine that, one sunny afternoon, a father takes his son to the art museum. There they are, and they stroll through the various rooms, and after a while they get to the sculpture exhibits. The little boy is especially attracted by one particular objet d'art in that room; he walks round and round this piece of sculpture, views it from different angles, bends down and reads the plaque that gives its title, and then looks up at his father and asks

"Daddy, is that really what a unicorn looks like?"

I submit to you that, as model assessors, our situation is much like that of this parent. We ask ourselves, for example, "Does this model really portray how the economy behaves?" Folks, there ain't no such thing as "the economy". There simply is no well-defined physical object or set of objects, clearly distinguished from all other objects and with behavior clearly distinguished from everything else that's going on in the world, which can serve as the referent for that question.

"The economy" is an extremely high-order intellectual construct, the details of the construction differing among economic scholars and philosophers. It is , like the unicorn, a myth -- an extraordinarily useful myth, but a myth nevertheless. To ask what a myth "really" is or does, is to ask a very difficult kind of question, and even understanding what we might mean by an "answer" poses a difficult task of conceptual analysis. (That is, the task of system identification, which is usually thought of as a necessary precursor to system analysis and prediction, is here more a matter of system definition and (mental) creation. Cf. Emery [21; pp. 43-44].) In like vein, to assess a model's ability to forecast "correctly" the interactions between the energy sector and the (rest of the) economy is to be concerned with predictions on the nature of the progeny spawned by the mating of two

myths, a project which might well daunt even the least inhibited of the medieval compilers of bestiaries.

In this context, I invite your attention to Hollaway's [38] identification of such down-to-earth phenomena as price and income levels as having normative rather than exclusively positive significance. And I cannot resist citing one pertinent though controversial passage from Kenneth Boulding's brilliant presidential lecture at the January 1980 annual meeting of the AAAS [7]: "A widespread illusion about science is that its basic theoretical images and paradigms are the result of inductive reasoning and experiments. It would be truer to say that science is the product of organized fantasy about the real world, tested constantly by an internal logic of necessity and an external public record of expectations, both realized and disappointed." Here is further corroboration from a distinguished bioscientist (de Duve [17; p. 3]): "As scientists we do not merely read the book of Nature. We write it -- that is the way our science progresses. But we must accept our concepts for what they are, provisional approximations that are as much fictions of our minds as they are faithful depictions of facts."

Obviously I do not mean, with all this talk of myths and fantasies and fictions, to imply that applied modeling is impossible or useless, any more than Boulding or de Duve would infer the fruitlessness of scientific activity. But I do suggest that hopes for better understanding of the modeling enterprise and its products can only be clouded, if a faulty viewpoint of naive Realism goes unchallenged.

3.3 The Elephant/Iceberg Problem

There is a whole nest of problems squirming here, which I will try to encapsulate in the following bit of pidgin-Zen: the elephant is only the tip of the iceberg. Let me explain.

When one sets out to appraise one of today's significant decision-aid models, the first property that is perfectly clear is that the thing is big, a veritable elephant of size and complexity to get hold of and understand. But we bravely say, "Well, this is our job," and proceed to plan somehow to assess this elephant. We will dissect it and check that its vital organs are properly connected, we will assay tne quality of the ivory in its tusks, we will observe whether it follows the mahout's instructions properly rather than running rogue in response to any unfamiliar type of goad, and so forth.

But as we really get into the job, we find that indeed the elephant, massive though it be, is only the tip of the iceberg that defines our true proper responsible area of concern. Another piece of that iceberg

(metaphors are about to be mixed into a fine froth) is the elephant's diet, the data on which it is fed. Typically a major part of that food is synthetic -- or, more forthrightly, concocted -- so that its nutritive value may be dubious (see, e.g. [24], Sections 6 and 7 of Freedman [22] together with Section V of Hopkins [39], also pp. 150-151 of [28]). Tracing such feed back through its intermediate processing stages to its raw ingredients, though clearly necessary for the independent-verification virtue of a "data audit", can be a formidable task (cf. Hirschfeld [36]). And I suspect that the particular concoction procedures employed (for estimation or aggregation or allocation or whatever) were not in general explicitly chosen to be optimal relative to the way those particular concoctions were going to be used in the latter stages of the modeling process. The development and use of data-transformation techniques deliberately tuned (e.g., in terms of "best-fit" criteria) to the error-cost structure of their products' intended use, seems to me an important technical area for modeling science.

Another chunk of the iceberg involves scenario formulation. Policy problems, in vivo, do not generally arise nicely formulated in terms of particular settings for the parameters of a model. If I would like to influence what answers a model gives, let me be the one who formulates for that model the scenario describing the decision problem at hand. Grant me that privilege, and (in the words of one participant in the discussion-session I attended) I can probably make that model dance to pretty well any tune that is desired. Cf. Kresge [46; p. 193]. The situation is, alas, quite different from that held up by Arthus [2, p. 20] as an ideal for scientific investigation: "The question raised must be such that it permits an answer, and that this answer may be clear and may if ever possible be expressed by yes or no. Ambiguous, approximate answers, full of reservations and double meaning are not conducive to progress and must be condemned But the question, of course, must have been put in a certain way. If a discussion remains confused it is not always because somebody answers in vague terms but often because the questions have been wrongly put."

But if the input channels to the model are too well-guarded for any flimflammery there, then let me be the one who interprets the model's outputs for the user, reading the entrails through appropriately tinted lenses, applying "subjective adjustments" in what Freedman [24] calls "the exercise of judgement through the back door," and standing well-poised to exploit what Arthus [2, p. 11] describes as the "contrast (of) the rigidity of the fact with the plasticity of its interpretation." And this area, that of the provisions for judgemental adjustment and communication and explication of model results, is yet another piece of the iceberg that requires exploration if the model's performance and merits in actual use as a decision-aid are to be well understood.

It has been extremely reassuring to me to see, in last year's Workshop and this one, increased and more sophisticated recognition of the need for extending assessment activity from just the elephant itself to the rest of the iceberg (e.g. Cazalet [12], Marcuse et al [49], Mayers [52], Hogan [37]). It is going to be terribly difficult, but it has to be done.

For example it is simply not enough, in my opinion, merely to assess a model as a mechanical object without regard for what the modelers' insights and expertise can contribute to its performance. (Cf. Greenberger et al [34; pp. 335-7], Ascher [3; pp. 64-70], Nissen [55; p. 272] and pp. 203-204 of [28]). To do so is like going down to the end of the block to look for the key you dropped, because the street lights are brighter there; it ignores the reality that the public and its representatives are (or should be) concerned for the quality of what flows from the entire model-use process.

But please don't misunderstand me; I said that such "depersonalized" analyses were by themselves insufficient, not that they were meaningless or (properly interpreted) misleading. In fact, I think they are necessary for an understanding of what the model per se does, and whether (as my wine-fancying friends might say) it is likely to travel well; cf. Lady [47; p. 10]. And there may be good reasons, such as those proposed by Kresge [45], why a part of such analysis should even be performed at a considerable remove from the modelers.

I agree with those (e.g. Mulvey [53; p. 188]) who find that modeling today is still largely an art form. (Cf. Klein's contrast [44; p. 9] of modern progress with the "artistic, subjective and personal" nature of earlier economic prediction.) Yet this element of artistry does not imply that no useful discussion of the product is possible without reference to the artist. For instance, musical scholars and musicologists can and do undertake technical analyses and aesthetic evaluations of Chopin's etudes despite having no recording of Chopin playing them. Part of what composing is all about is the creation of musical works that will continue to display beauty and give pleasure when performed by others, in different places and at different times. (It could prove interesting to attempt a "contrast and compare" discussion of the respective roles of a composer's immediate patron and a modeler's immediate client.)

The case of Paganini might be raised as a possible counterexample. It would indeed be especially difficult to make any judgements about his compositions without the maestro himself at hand to perform them. But Paganini was composing not for the sake of the ideals described above, but rather to provide showcases for his own virtuostic talents. I don't think the leading modelers today are operating under that particular

motivation, though the intrinsic fascination of the modeling game creates some real temptations for the ego (cf. Massarik [51; pp. 16-18] on "The Model Builder's Culture"). At any rate, I continue to ascribe value to the traditional scientific criteria of reproducibility and portability, while acknowledging that full-scale assessment must extend beyond these properties of the model to include the human elements of the modeling/analysis system.

In their paper at last year's workshop, Peter House and Roger Ball [40; p. 159] assert that while the scientist can "punt" when he or she comes to the end of what the available models and analytical methodologies say, the (decision) analyst cannot. To which Roger Glassey in the ensuing discussion [28; p. 167] rejoined, "The modeler, when he reaches the end of the limits of his science, at that point should confess that he has come to the end of his knowledge as a modeler, and then fall silent." (Possible second thoughts appear in the "--we must respond, with or without a model" of Glassey and Greenberg [31].

Surely, whether the analyst can or can't (or, must or musn't) punt, depends on what game is being played. If, like a geometer trying to construct the regular n-gon using straight-edge and compasses only, the modeler is only permitted to communicate that which has been established by "a scientific method" -- presumably the formalized reproducable approach of or akin to the model itself -- then indeed, after this communication is made, the rest must be silence. And if, as Glassey (op. cit.) explicitly opines, the modeler really has nothing more to say that is worth hearing, then such a restriction is prudent. But in the absence of experimental evidence (perhaps well worth pursuing), I am dubious of the "nothing else worth hearing" hypothesis. It seems more likely to me that the analyst is bright, has built up a highly trained intuition in the course of working and playing with the model and its data, and should not be forbidden to contribute the benefits of that informed though unformalized intuition to the cogitations of the decisionmaker.

What is essential to maintain is truth in labeling. And so these "extracurricular" contributions need to be labeled explicitly as outputs of the modeler's intuition, not of the model itself. Professional ethics demand that the modeler make this distinction very clear to the client.

Now a more delicate point arises. If the decisionmaker, in presenting recommendations derived in part from such intuition-based input, clouds over the distinction, what is the modeler to do? If the witness giving testimony before Congress says in effect, like the principal in an E.F. Hutton commercial, "My model says ---," when in fact the modeler operating in intuitive mode said it -- or even worse, the model "said it" in the sense of a confession extorted in the police station's back room, with its inputs twisted and its logic "adjusted" to produce a desired result -- if this is what's going on, does the modeler have a professional

responsibility to blow the whistle and try to set the record straight? My own answer is "yes". It is an unpleasant question, that I thought needed ventilation.

4. Credibility in an Age of Mistrust

The currently popular term "credibility" seems to me a rather good choice for expressing, in a single word, that property of a model or model-application at which assessment for decision-aid capability is aimed. But it's subject to serious ambiguity if not qualified with care, and it is all too easy -- speaking from one's own institutional setting, own viewpoint -- to skimp those qualifications, thereby confusing or misleading listeners who don't share the same preconceptions.

For example, does "credibility" refer to the degree to which belief or confidence _is_ elicited, or the degree to which it _should_ be elicited in an ideal world with an ideally well-informed and open-minded population of potential believers?

Credibility to _whom_? An immediate client? The Congress? A wider public? To assessors? That important distinctions hinge on the answer has been noted by prior speakers in this symposium; see also Gass and Joel [30].

Credibility of _what_? Of a model? An entire modeling system? Particular applications of such a system? "Analytical results"? [31]. We really need to be careful both to specify which of these we mean at a given moment, and to keep the full range of possibilities in mind (cf. Joel and Maybee [43]) as we work towards methods for estimating and enhancing credibility.

Churchman [14], discussing the assessment and validation of scientific theories, observes that "a whole set of problems as to measures of relevance and confirmation must be considered," and goes on to pose sharply one of the basic issues: "Some, like Carnap, believe the degree of confirmation of a hypothesis can be expressed independent of the use to which a hypothesis is to be put. For them, the degree of confirmation expresses merely a relationship between the set of observation sentences and the hypothesis. Others, like Wald and many statistical decision theorists, believe that a weight function which helps to define the seriousness of a mistake is essential."

It seems to me that Wald's stance, though intrinsically more difficult to implement because of its requirement for a weight function, is

the more appropriate in considering the reliability of a model in a particular decision context. Going beyond the limits of my own technical competence, I would urge that in grappling with the terribly difficult conceptual problems of credibility (cf. House and Ball, [40]), we not neglect the intellectual resources that may be available from the scholarly community most closely involved with such questions: the logicians and philosophers of science specializing in the area described by such terms as "inductive logic" and "confirmation theory." Some of this work goes further, explicitly linking degree-of-belief with decision and choice. The "classical names" in the field are those of Carnap and Reichenbach; its nature can be sampled by examining such works as Burks [8], Carnap and Jeffrey [10], and Shafer [59].

When credibility is being spoken of, just what is it that someone is supposed to be believing about the model? That particular numerical outputs are right on the head? Surely not. As one discussant at an earlier panel said, "One thing we know about the numbers that come out, is that they're wrong." It is inconceivable that they would be literally, exactly correct. Accurate enough, with high enough confidence -- that seems more like it. But how much is "enough"?

My own feeling, about what it is that ought to be believed, is this: that use of this particular model or modeling approach, this particular system, gives the decision-maker a good chance -- a better chance than the available alternatives -- of correctly identifying the better policy or policies among the range of accessable alternatives. And I would like to see at least part of our assessment and assessment-method development programs aimed explicitly and directly at this sort of functional ordinal-flavored criterion, rather than at intermediate cardinal-flavored desiderata.

Under some of its interpretations, credibility can depend a great deal on the amount of credulity, or even just plain trust, that's around. It seems a rather scarce commodity today, partly because of healthy skepticism by a more alert and better-informed public, and partly for uglier reasons that I'll touch on later. This is one of the reasons, I think, why the originators of some of the leading models find themselves suffering under a particular and cruel kind of burden.

Let me speak, for example, of PIES. The operations research community is forever and with good reason bemoaning the infrequency with which its contributions achieve relevance, responsiveness, end use, and significant impact. The PIES makers, operating under extraordinarily stressful conditions,performed mighty feats of analysis, intellectual integration, model creation, and persuasiveness, and succeeded admirably at achieving many of these goals that the O.R. profession sets up as lofty standards. But instead of being awarded medals they find themselves, in the context of "assessment", being berated for failing to satisfy

criteria they feel are irrelevant, ultra purist, and the like. (See, e.g., the discussions interpersed in [33].) It is not unnatural, in such circumstances, for a certain bitterness to result. We have heard it surface, quite unmistakably, at this symposium and its predecessors.

Some of it arises because the fault-finding is (or is perceived as) directed at the individuals involved, rather than the situation or institutional setting in which the work was done. The closest analogy -- and it isn't a terribly good one -- is with the responses to returning Vietnam-war veterans by those revolted by or disgusted with that conflict, and with those veterans' reaction to these responses.

But there is another point to be made. Sure, it's terrific if an analyst working hard and smart, or an analysis team or a modeling team, can give decision-makers very good advice and insights about hard problems. That by itself used to be the name of the game, but the game has changed over the last decade or so, in what I call "an age of mistrust."

It is fact rather than opinion that, during this period, the public has been betrayed by holders of high office and esteem in the governmental sector, the private sector, and, some feel, in the professions. Moreover, the informed public knows this in many cases, and suspects it in many more (probably not always the right ones). That public, for darned good reasons, will no longer give ready trust or even grudging trust to pronouncements by authority figures. (Cf. Cantlon's reference [9] to "the public's post-Watergate, post-Vietnam sag in confidence about national institutions -- and decisions.")

It follows that if the analyst-modeler can not only provide good advice for the client, the decisionmaker, but can also provide means for demonstrating to a concerned and skeptical <u>public</u> that this advice is indeed well-founded -- or, better still, provide tools which that public can apply <u>itself</u> to verify that the advice is good -- then that analyst-modeler has performed an even more notable, extraordinary and praiseworthy feat. I submit that the requirements for laurel wreaths, in this suspecting era, have been raised to the level of accomplishment just described, one we are still a very long way from being able to attain.

(This course of events has historical parallels, as in Beth's account [6; p. 58] of the origin of formal logic: "It arose out of the needs of argumentation. In the oldest thinkers we never find a deductively coherent argumentation; they restrict themselves to a dogmatic exposition of their views. Formal reasoning only occurs when the need is felt to refute an opponent or to defend one's own views against the views of others.")

A mundane but very real contributor to generalized distrust, is widespread disappointment over the quality and performance of many consumer products. Some desultory reading [16, 54] on the "theory" of fraud, in consumer goods and services, turned up an interesting three-way classification of certain qualities of such products.

One class consists of "search" qualities, those that can be ascertained without purchase. For example, the style of a suit -- I can look at the suit in the store or store-window and make a reasonably good judgement on how well I'd like it. Then there's what's called "experience" qualities, not discoverable by mere inspection, but only through purchase and use -- e.g., the taste of a can of tunafish. The third and last category consists of "credence" properties. They may be very important, but are not revealed through normal use of the product; rather, they can be assessed only by costly information-gathering perhaps involving the consultation of experts. For example, whether or not you really needed that expensive auto repair, or to have your appendix taken out.

Of course model products are not consumer goods, and I hope we will never need to be talking about "modeling fraud" (cf., however, the dispute [11] over ballistic-missile defense analysis). But it still seems to me that this trichotomy provides one useful analytical framework for thinking about why model assessment is needed, why calls for it have arisen, and what forms and gradations it may take.

5. Decisions: Responsibilities and Strategies

I spoke earlier about the need to develop techniques, for model analysis, which explicitly reflect the context of the decision processes in which those models are supposed to be useful. Now I'd like to say just a bit more about that context, beginning with the following quotation (Young [70; p. 2]):

"It has become accepted that decision-makers, especially in government, have a duty to cope with changes and control them. This they can hardly do unless they make an attempt to foresee them. This is of course an extremely hazardous undertaking since it is -- part of the definition of the future that it is unknown. The job becomes more difficult -- and more necessary -- just because men have gained increasing control over their environment."

It is exactly this difficult and necessary job that the modeler seeks to assist. In past centuries, when technologies and social institutions

changed much more slowly, the principal uncertainties and hazards of life involved famine, illness and the like -- things that operated on a personal level or had "Act of God" status, so that one couldn't very well expect government to do much about them. (The two principal exceptions -- wars and taxes -- were in fact major causes of public unrest.) Today we are longer-lived; harvests are more predictable and crop failures less disastrous -- the vicissitudes of life arise less at the individual level and more at the regional and national scale where the foresight and wisdom of government leaders can make a difference. But at the same time, "because men have gained increasing control over their environment," that foresight must extend beyond the forces of Nature to what other decisionmakers, in our own nation and in others, may do. Such contingencies constitute significant "singularities of uncertainty" poorly suited to statistical treatment (cf. Freedman [23], Bassie [4; p. 8]), so that it is hard to suggest alternatives -- but see Emery [21] -- to oracular (i.e., Delphi) techniques. (It will be interesting to follow the fate of Ascher's suggestion [3; p. 211] of specialized "surprise-sensitive forecasting" to combat the problem of "assumption drag"; re the latter, cf. Simon [61].)

The easy choices are rarely even dignified with the term "decision". That's the real point of Ron Howard's statement [41] in a recent issue of Operations Research, that "Decision making is what you do when you don't know what to do." There are many choice situations in which your previously-developed ethical or religious or ideological system delivers a snap answer; having principles is -- among other things -- an invaluable device for efficiency of choice. There are other cases where the issues are trivial or simple enough, and you're expert enough, that you can settle the matter right off without even a conscious act of "deciding". So when the phrase "decision making" does arise, you can be pretty sure the question involved is both significant and perplexing.

That "part of the definition of the future is that it is unknown" is, I think, the fundamental point that tortures many of us concerned with modeling and model assessment. In trying to develop validation concepts and assessment techniques, our minds are deeply colored by past immersion in the value schemes of the sciences, with their incredible predictive power utilizing "the unreasonable effectiveness of mathematics" (Wigner [67]). And so we keep trying to find some way of strictly holding decision models, policy-aid models, to standards comparable to those traditional for scientific theories. And when the pain finally drives us to recognize the futility of such insistence, we lurch to the opposite extreme and declaim that rigor and notions of "scientific method" make no sense whatever in such appraisals. It is basically the unknowability of the future that gnaws at our vitals and inflicts the emotional and philosophical tensions that send us careening between such irrational extremes.

You'll recall that Larry Mayer, in his talk earlier in this symposium, offered us a theorem: all models are optimal. In the same spirit of hyperbole, I offer a countertheorem: all models are crummy.

The proof is very short. What we want, what we desperately long for, what we genuinely need, is an ability successfully to forecast well into the future the course of the complex and uncertain matters on which policy decisions are required. Relative to this heart's desire, no model comes anywhere near to measuring up, and since the future is unknowable, most likely none possibly can. (Is there an axiom scheme, like some of those for quantum mechanics, in which this can be formally demonstrated?) Ergo, all models are crummy.

But still, we must try to develop some ways to distinguish degrees of crumminess, to help the decisionmaker (and the citizen) assess the relative utilities of the various modeling tools that might be present on the stage. And -- once again -- I think that more of that effort should involve explicit consideration of the model's context in a decision or policy process. One source of useful ideas about such processes is the writings of Charles Lindblom, e.g. [48].

I suspect that many of my contemporaries in Operations Research were brought up, as I was, to view Lindblom as a bete noire. He was accused of being the champion of "muddling through" -- the champion not of the sort of bold, rationally-derived optimal decisions we were taught to revere, but rather of a cowardly and unesthetic policy of incremental change. But as we went on and encountered a bit more of the world, and found out through personal experience a little more about what responsible decision-making is like, we may have become rather more sympathetic to what he had to say.

One of his most striking points, if I may paraphrase it loosely in modern lingo, is that significant decision problems are NP-hard. Full execution of the traditional paradigm of rational decision-making is not just difficult, it's impossible. It is literally never possible to carry out all of that paradigm's prescriptions. Here is the exact passage [48; p. 6]:

"For a complex problem it is never possible to be clear about all the values affected by a decision, including their relative weights, never possible in fact to ascertain all possible means of achieving the objective, never possible exhaustively to trace through the chain of consequences that are attached to each possible means."

One response to these realities, the traditional one, is to recognize the unattainability of the ideal decision process but to seek to approximate that ideal as well as possible under limitations of time, information and resources. A quite different approach, which Lindblom espouses and claims to observe in the behavior of superior decision-makers, he calls "strategic problem solving."

Again I quote, "The decision maker has to acknowledge that he must take short cuts, must leave important aspects of his problem out of his analysis, must make judgements on the basis of values only roughly perceived, and must make do with dodges and stratagems that are not scientifically respectable -- He can out corners, omit, improvise amateurishly and intuitively. Or he can try to develop a well studied and thought out strategy to guide him in his shortcuts, omissions and dodges. (The strategic analyst) chooses the latter. He takes seriously the need for developing a strategy for coping with problems on which conventional scientific prescriptions give him inadequate guidance."

Lindblom goes on to identify a number of elements of such a strategy. One of them is indeed incrementalism; incremental changes are typically those for which you have enough information to do some sensible analysis. Also, they are likely to be the changes most likely of acceptance and implementation, so that more rapid progress may be possible through actual execution of a sequence of such changes than through opting for the bold and radical new step.

A second guideline is to stay reversible, or at least flexible; more information will be gained as time proceeds, and the decision-maker should try to preserve the ability to make revisions in the light of that information. This strikes a very responsive chord in me; I think a major theoretical shortcoming in most of today's policy-aid models is that their basic framework is that of "the good decision, now", rather than one of developing good decision rules that explicitly take into account the availability of further information as events unfold, and of giving guidance on the proper balance between (i) full commitment and (ii) hanging loose so as to be better able to exploit what will be learned later. (For a somewhat dated sampler on these themes of decision-rules and "stochastic programming with recourse," see, e.g., Charnes and Kirby [13], Sengupta [58], Theil [64], Walkup and Wets [65].)

A third of Lindblom's maxims (I will not list them all) is: Be remedial! That is, if the task of gaining consensus on desirable values and ends proves unmanageable (as well it may), it may prove easier to secure agreement as to what in present trends promises undesirable consequences and what can be done to avoid them -- trying "to define not the goal to be sought but the situation from which escape is desired." This idea of negative steering seems to me a very interesting one in our picture of the policy process as the milieu for model application. I was quite sorry to see it in Lindblom's list, because I'd thought of it on my own some years ago and had been rather proud of myself for doing so. But that's life.

To return to the point which prompted this excursion: if those who study public-sector decisionmakers have observed significant regularities in the processes they follow, should not our model-assessment activities evolve to take more direct and explicit account of those regularities than is now done?

6. Load-Sharing and MODELWORLD

Most of what I've said this morning, most of what we've discussed during this symposium, is by no means peculiar to the field of energy. It's natural to consider why it's been the Department of Energy which, in an admirable exercise in masochism, has made available the substantial funding and internal effort that have been invested both in assessment of its own models and in advancing the general art of model analysis.

One might, for example, note the unique role of PIES in energy policy deliberations at a crucial time. One might recall the particular skepticism and hostility displaced from the policy recommendations to the supporting models themselves. There are several first-hand accounts of this background (e.g., Nissen [55]), so I needn't belabor it. But though aware of the many factors uncongenial to evaluation and analysis efforts in the public sector (cf. Dye [18; pp. 97-98, 103-107], Greenberger et al [34]), I am disappointed that more departments of the government are not developing comparable pangs of conscience and beginning comparable efforts in the assessment of their own considerable modeling activities.

I have heard it said that modeling is more critical in the energy policy area because of the big investments and long lead times involved. But surely much the same can be said for transportation: when you pour concrete for highways, that concrete is going to be there for a long time. In a way the situation is even worse: you cannot very well transfer highway capacity from one region or corridor to another if you have misestimated where the big jump in demand is going to be, but the corresponding problem for energy can be accomodated at least in part through use of fuel shipments and energy transmission grids.

Perhaps later workshops in this series will, through their speakers and invited attendees, be able to promote a wider exchange of information about -- and hopefully, a wider spread of activity in -- model assessment work in agencies outside DOE. Energy is carrying a considerable load of investment and nose-bloodying in an effort that's of value to a much larger community, and there really ought to be more load-sharing.

In particular, there are some more basic kinds of research that I think should be attempted. Efforts to understand better how to do modeling, how to assess modeling, necessarily suffer somewhat by being thrust immediately into the battleground of the very largest models operating on the most controversial policy problems. Sure, that's the way to attract a lot of interest right off. It's the context in which funds are most likely to be available from mission-oriented agencies. But it does <u>not</u> represent all that's needed for an orderly sustained development of basic knowledge. Modeling science is in many ways still in its infancy, and a battleground is not the ideal environment for an infant.

Looking only or primarily at the biggest models dealing with the hottest issues is a little like starting, in physics, with the nuclear reactor rather than the simple pendulum. So my mottos, for now, are "Small is informative!" and "Cool it!" I suggest that as part of our efforts to learn about the effectiveness (in terms of predictive quality and usefulness to decisionmakers) of alternative approaches in modeling, there is need for a line of systematic research that starts with simpler situations.

It would begin with historical and synthetic studies of smaller modeling systems dealing with issues which are "cool" at present. (Ascher [3] and Greenberger _et al_ [34] can be viewed, in part, as contributions to the historical branch of this work.) It would develop a taxonomy of alternative ways of planning and conducting model development projects and model applications. It would attempt to develop theories to predict in which kinds of settings which of these alternatives are preferable. And it would proceed to test these theories, initially by small-scale experiments and then if they prove promising, by a larger kind of experimental activity.

Let me refer to that larger activity or facility as _Modelworld_. This is an idea stimulated by the RAND Corporation's former Logistics Systems Laboratory, which did some very notable work. The essential notion is to develop a synthetic little world with which to challenge the would-be modeler or modeling approach. It would include random processes, perhaps some human interventions, but all of them controlled, recorded, and reproducible. It would include information instruments analogous to the Census and other common or special real-world data sources. And in that controlled environment one would experiment with different approaches to modeling efforts aimed at learning about how that synthetic world operates and (most important) how to make good decisions about "policy" questions posed in its context. One would try to determine how these approaches compare under different settings of the experimental factors: time limitations for the modeling effort, resource constraints, quality of the data probes and measurement devices available to the modeler for exploring Modelworld, complexity of the underlying problem, etc. A natural test-bed would be at hand for the previously described research on the development of functional documentation standards.

Modelworld would not be a trivial or inexpensive undertaking. But I think that something like it is, quite literally, indispensable to the growth of a modeling science.

7. Concluding Remarks

Most of my comments today were prepared prior to this meeting. I am not equal to the task of quickly synthesizing and reacting to the richness of the intellectual stimuli that have bombarded me during these three days.

I did, however, note with interest two particular remarks made by two of our distinguished colleagues from the EIA. George Lady, speaking about the need for measures of credibility of DOE's in-use modeling systems, described that need as a requirement for "quality control for this aircraft which is in flight right now." And Harvey Greenberg, looking back to the good old days, said (I have added one piece of underlining), "Much of our best work was done in a crash mode."

Perhaps the apposition of those two remarks speaks for itself, although I'm not totally sure what it's saying.

At the Operations Research Society meeting just a couple of weeks ago, Dave Wood gave a paper [69] in which he said, "We all know that in order to aid decisions, models have to leap beyond the data, and one of the difficult issues in assessment is judging whether or not the leaps are artfully made." From this I conclude -- and concur -- that the assessor has to be somewhat of a ballet critic.

Combining this idea with the previous ones, I get a picture of a giant airplane operating in crash mode, over a sea of icebergs, with unicorns and elephants milling about in the cargo hold, and with the pilot doing jetes and pirouettes in the control room. One is certainly tempted to prefer the train. But for this trip, there is no train.

It's risky to invite someone to present his "reflections" on a topic about which he's been reflecting for a long time. This has been a long talk, but the end is near. To provide a further target for your reflections, I'd like to go back to the 1961 War Gaming Symposium mentioned earlier, and to recall the discussion of "success" offered by one of its speakers, Hebron Adams [1; pp. 52-53]. (I have changed some language from "game" to "model" to better fit the present setting.)

"Despite appearances, a model is not successful just because it provides technological employment for modelers and programmers. It is not necessarily successful because it has gained customer acceptance (to date, few have), and it is not necessarily a failure because it has not. It is most emphatically not successful simply because it has seconded expert opinion.

"A model is partially successful if it extends human knowledge a little, if it is right when the experts are wrong. It will never be self-evidently

right and it may never be completely right. But if, after unbiased examination of the course of a model run and after further analysis prompted by its output, expert opinion is changed in whole or in part, the model is partially successful.

"A model is completely successful only if it is right. With war games, one never knows until the shooting is over and the real world's results can be compared with the synthetic world's predictions. If the model is a failure, the results may be disastrous. That is the shadow under which the modeler must always work."

And finally, I thought I would close with words more eloquent than any I myself could muster, to remind us once again of the limitations of modeling applied to a world exhibiting the blessing and curse of ever-surprising change, and of how those limitations make the job of model assessment so difficult, so treacherous, yet so important. The words are those of T.S. Eliot in one of his Four Quartets ([19; p. 10], see also the opening of [20]):

> "There is, it seems to us,
> At best, only a limited value
> In the knowledge derived from experience.
> The knowledge imposes a pattern, and falsifies,
> For the pattern is new in every moment
> And every moment is a new and shocking
> Valuation of all we have been."

Thank you, Mr. Eliot. Thank you, Symposium sponsors and organizers and hosts. And thank you, audience, for your attention and patience.

8. References

1. H. E. Adams, "Machine-Played Games," pp. 35-53 in [56].

2. M. Arthus, Philosophy of Scientific Investigation (H. E. Sigerist, trans.), The Johns Hopkins Press, Baltimore, 1943.

3. W. Ascher, Forecasting: An Appraisal for Policy-Makers and Planners, Johns Hopkins University Press, Baltimore, 1978.

4. V. L. Bassie, Uncertainty in Forecasting and Policy Formation, 1958-1959 J. Anderson Fitzgerald Lecture in the Dynamics of Business Enterprise, University of Texas, Austin, 1959.

5. M. L. Baughman, "Reflections on the Model Assessment Process: A Modeler's Perspective," pp. 197-210 in [28].

6. E. W. Beth, Science, A Road to Wisdom: Collected Philosophical Essays, D. Reidel, Dordrech, Holland, 1968.

7. K. E. Boulding, "Science: Our Common Heritage," Science 207 (22 February 1980), 831-836.

8. A. W. Burks, Chance, Choice, Reason, University of Chicago Press, Chicago, 1977.

9. J. Cantlon, "Summary of the Workshop," in Long-Range Environmental Outlook, National Academy of Sciences, Washington, D.C., 1980.

10. R. Carnap and R. C. Jeffrey, Studies in Inductive Logic and Probability (vol. 1), University of California Press, Berkeley, 1971.

11. T. E. Caywood et al, Guidelines for the Practice of Operations Research, Oper. Res. 19 (1971), entire No. 5.

12. E. G. Cazalet, "A Decision Analyst's View of Model Assessment," pp. 325-326 in [28].

13. A. Charnes and M. Kirby, "Optimal Decision Rules for the E-Model of Chance-Constrained Programming," Cah. Centre Etudes Rech. Oper. 8 (1966), 5-44.

14. C. W. Churchman, "Towards a Mathematics of Social Science," pp. 29-36 in [50].

15. A. E. Daniels, "The User's Viewpoint on the Creation and Use of a Large Simulation Model," pp. 84-96 in [56].

16. M. R. Darby and E. Karni, "Free Competition and the Optimal Amount of Fraud," J. Law and Economics 16 (1973), 67-88.

17. C. de Duve, "The Lysosome in Retrospect," pp. 3-40 in Lysosomes in Biology and Pathology (vol. 1), J. T. Dingle and H. B. Fell (ed.), North-Holland, Amsterdam and London, 1969.

18. T. R. Dye, Policy Analysis, University of Alabama Press, 1976.

19. T. S. Eliot, East Coker, Faber and Faber, London, 1940.

20. T. S. Eliot, Burnt Norton, Faber and Faber, London, 1943.

21. F. E. Emery, "Concepts, Methods and Anticipations," pp. 41-70 in [70].

22. D. Freedman, "Assessment of the READ Model," pp. 365-395 of [28].

23. D. Freedman, "Uncertainties in Model Forecasts," working paper, September 1979.

24. D. Freedman, "Are Energy Models Credible?", these Proceedings.

25. S. I. Gass, Computer Model Documentation: A Review and an Approach, National Bureau of Standards Special Publication 500-39, Washington, D.C., February 1979.

26. S. I. Gass (ed.), Utility and Use of Large-Scale Mathematical Models, National Bureau of Standards Special Publication 534, Washington, D.C., May 1979.

27. S. I. Gass et al, Interim Report on Model Assessment Methodology: Documentation Assessment, National Bureau of Standards Technical Report NBSIR 80-1971, January 1980.

28. S. I. Gass (ed.), Validation and Assessment Issues of Energy Models, National Bureau of Standards Special Publication 569, Washington, D.C., February 1980.

29. S. I. Gass, "Assessing Ways to Improve the Utility of Large-Scale Models," pp. 273-254 in [28].

30. S. I. Gass and L. S. Joel, Concepts of Model Confidence, National Bureau of Standards Technical Report NBSIR 80-2053, June 1980.

31. C. R. Glassey and H. J. Greenberg, "Model Credibility: How Much Does it Depend on Data Quality?", these Proceedings.

32. A. J. Goldman, "Operations Research Research and Governmental O.R.," pp. 13-18 in Operations Research: Proceedings of a Conference for Washington Area Government Agencies, J. A. Joseph (ed.), National Bureau of Standards Miscellaneous Publication 294, Washington, D.C., December 1967.

33. H. Greenberg, "The FEA Project Independence Experiment," pp. 111-222 in [26].

34. M. Greenberger, M. A. Crenson and B. L. Crissey, Models in the Policy Process, Russell Sage Foundation, New York, 1976.

35. H. Guetzkow et al, Simulation in International Relations: Developments for Research and Teaching, Prentice-Hall, Englewood Cliffs, N.J., 1963.

36. D. S. Hirschfeld, Investigation of Underlying Data: Midterm Oil and Gas Supply Modeling System, Report No. KFR245-79 (to National Bureau of Standards), Ketron, Inc., Arlington, VA, January 1980.

37. W. W. Hogan, "Discussant's Remarks," pp. 63-64 in [28].

38. M. L. Hollaway, "Issue Paper on Model Standards to Aid Assessments," this Symposium.

39. F. Hopkins, "A Modeler's View of the READ Model Assessment Process," pp. 397-429 in [28].

40. P. W. House and R. H. Ball, "Validation: A Modern Day Snipe Hunt? Conceptual Difficulties of Validating Models," pp. 153-165 in [28].

41. R. A. Howard, "An Assessment of Decision Analysis," Oper. Res. 28 (1980), 4-27.

42. S. W. Huck and H. M. Sandler, Rival Hypotheses: Alternative Interpretations of Data Based Conclusions, Harper and Row, New York, 1979.

43. L. S. Joel and J. S. Maybee, "Model Confidence," this Symposium.

44. L. R. Klein, An Essay on the Theory of Economic Prediction, Markham, Chicago, 1971.

45. D. Kresge, "The EPRI/NBER Energy Model Assessment Project," pp. 123-136 in [26].

46. D. Kresge, "An Approach to Independent Model Assessment," pp. 183-196 of [28].

47. G. L. Lady, "Model Assessment and Validation: Issues, Structures, and Energy Information Administration Program Goals," pp. 5-22 in [28].

48. C. E. Lindblom, Strategies for Decision Making, Thirty-fifth Edmund J. James Lecture on Government, University of Illinois, 1972.

49. W. Marcuse et al., "Validation Issues - A View from the Trenches," pp. 337-353 in [28].

50. F. Massarik and P. Ratoosh (ed.), Mathematical Explorations in Behavioral Science, Richard D. Irwin, Inc. and the Dorsey Press, Homewood, IL, 1965.

51. F. Massarik, "Magic, Models, Man and the Cultures of Mathematics," pp. 7-21 of [50].

52. L. S. Mayer , "A Perspective for Energy Model Validation," pp. 477-496 in [28].

53. J. Mulvey, "Strategies in Model Management," pp. 177-194 in [26].

54. P. Nelson, "Information and Consumer Behavior," J. Pol. Econ. 78 (1970), 311-329.

55. D. Nissen, "The Impact of Assessment on the Modeling Process," pp. 267-283 in [28].

56. J. Overholt (ed.), First War Gaming Symposium Proceedings, Washington Operations Research Council, Washington, D.C., 1962.

57. R. Richels, "Third Party Model Assessment: A Sponsor's Perspective," pp. 171-181 in [28].

58. J. K. Sengupta, Stochastic Programming: Methods and Applications, North-Holland and Elsevier, Amsterdam and New York, 1972.

59. G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, Princeton, 1976.

60. M. E. Shaw, "Model Access and Documentation," pp. 355-363 in [28].

61. J. L. Simon, "Resources, Population, Environment: An Oversupply of False Bad News," Science 208 (27 June 1980), 1431-1437.

62. R. M. Smullyan, What is the Name of This Book? The Riddle of Dracula and Other Logical Puzzles, Prentice-Hall, Englewood Cliffs, NJ, 1978.

63. S. L. Solomon, "Building Modelers: Teaching the Art of Simulation," Interfaces 10 (1980), 65-72.

64. H. Theil, Optimal Decision Rules for Government and Industry, North-Holland and Rand-McNally, Amsterdam and Chicago, 1968.

65. D. W. Walkup and R. J. B. Wets, "Stochastic Programming with Recourse," SIAM J. Appl. Math. 15 (1967), 1299-1314.

66. J. O. Wallace, "The Practical Meaning of Library Standards" and "The History and Philosophy of Library Standards," pp. 31-56 in Quantitative Methods in Librarianship: Standards, Research, Measurement, I. B. Hoadley and A. S. Clark (ed.), Greenwood Press, Westport, CN, 1972.

67. E. P. Wigner, "The Unreasonable Effectiveness of Mathematics in the Natural Sciences," Comm. Pure and Appl. Math. 13 (1960), 1-14.

68. D. O. Wood, "Model Assessment and the Policy Research Process: Current Practice and Future Promise," pp. 23-62 in [28].

69. D. O. Wood, "Energy Policy Model Evaluation: A Summary of Methods and Current Practice," Paper WD01.5, Abstract in TIMS/ORSA Bulletin No. 9, Joint National Meeting, Washington, D.C., May 5-7, 1980.

70. M. Young (ed.), Forecasting and the Social Sciences, Social Science Research Council, Heinemann Educational Books Ltd., London, 1968.

PROGRAM

DOE/NBS SYMPOSIUM

VALIDATION AND ASSESSMENT OF ENERGY MODELS

May 19-20-21, 1980
Administration Building
National Bureau of Standards
Gaithersburg, MD

Monday, May 19, 1980--Green Auditorium

9:00 a.m. - 9:15 a.m.:

WELCOME..Saul I. Gass, U. of MD/NBS
Symposium Chairman

Lincoln E. Moses
Administrator, EIA/DOE

Burton H. Colvin
Director, CAM/NBS

9:15 a.m. - 12:00 p.m.:

MODEL ANALYSIS--1980 AND BEYOND

Speakers: "Energy Model Evaluation and Analysis: Current Practice"......David O. Wood, MIT

"Humanizing Policy Analysis: Confronting the Paradox in Energy Policy Modeling".......Martin Greenberger, JHU

Discussants: Ken Hoffman, Mathtech
George Lady, DOE
James Sweeney, EMF

12:00 p.m. - 3:30 p.m.: (Lunch will be from 1:15 - 2:00 in the Main Cafeteria)

THE EIA MODEL VALIDATION PROGRAM

TOPIC: This session will review EIA's activities in evaluating the energy system projections and analyses published in the EIA Annual Report to Congress 1978, Volume III (ARC 1978). The speakers will critique the reports produced by the individuals and organizations represented on the panel.

Speakers: Richard Richels, EPRI
Jerry Hausman, MIT
Roy Shanker, Resource Planning

Panel: David Freedman, UCal at Berkeley
Richard Jackson, NBS
John Maybee, UCol
Alan Soyster, VPI&SU
David Wood, MIT

3:30 p.m. - 6:00 p.m.:

MODEL STRUCTURE AND DATA

Speakers: "Are Energy Models Credible?".......David Freedman, UCal at Berkeley
"Model Credibility: How Much Does it Depend on Data Quality?".........Roger Glassey, DOE and Harvey Greenberg, DOE

Discussants: Douglas Hale, DOE
Lawrence Mayer, UPenn
David Nissen, Chase Manhattan
David Pilati, Brookhaven

Tuesday, May 20, 1980

WORKSHOP DAY

There are five workshops as noted below. Each registrant of the Symposium is requested to sign up and participate in only one workshop. Workshop sign-up sheets will be available on May 19, along with the issue papers that describe the five workshop topic areas. These papers review the current state-of-the-art of an area and questions to be resolved. The purpose of a workshop is to obtain a consensus, if possible, of each topic's open issues and recommendations for future research. The chairmen will present a workshop summary on May 21, and prepare a report for the proceedings. The only way these workshops can be a success is by your active participation. The workshops are an all-day affair and are the only activities scheduled. They afford each participant an opportunity to meet other researchers, to influence future research, and to discuss their research and ideas in an informal and congenial forum.

WORKSHOP 1 -- Validating Composite Models
Chairmen: Andy Kydes, Brookhaven
Lewis Rubin, EPRI

WORKSHOP 2 -- The Measurement of Model Confidence
Chairmen: Lambert Joel, NBS
John Maybee, UCol

WORKSHOP 3 -- Model Standards to Aid Assessment
Chairman: Milton Holloway, TEAC

WORKSHOP 4 -- Sensitivity and Statistical Analysis of Models
Chairmen: Carl Harris, Ctr. for Mgmt. & Policy Research
Dave Hirshfeld, DSH Assoc.

WORKSHOP 5 -- Model Assessment Methodologies
Chairmen: James Gruhl, MIT
Thomas Woods, U. S. GAO

All workshops will be held in the Administration Building--Lecture Rooms A and B, Dining Room C, 10th Floor Conference Room, and the Green Auditorium. Room assignments will be posted and announced late Monday, May 19.

Wednesday, May 21, 1980

9:00 a.m. - 12:00 p.m.:

WORKSHOP REPORTS AND DISCUSSIONS

1. Validating Composite Models
2. The Measurement of Model Confidence
3. Model Standards to Aid Assessment
4. Sensitivity and Statistical Analysis of Models
5. Model Assessment Methodologies

12:00 p.m. - 1:00 p.m.:

FINAL SESSION

Speaker: "Reflections on Modeling"..............Alan J. Goldman, JHU/NBS

Note: There is no registration fee. There will be no scheduled coffee breaks; coffee may be obtained in the main cafeteria from 9 - 10:30 a.m. and 2 - 3:30 p.m. Breakfast is served until 8:30 a.m.; lunch is served from 11:30 a.m. to 1:30 p.m. The employees' lounge, across from the cafeteria, is available each day for informal discussions.

SYMPOSIUM PARTICIPANTS

Susan R. Abbasi
Environment & Natural Resources
Division, Library of Congress
425 Madison Bldg.
Washington, D. C. 20540

Gary B. Ackerman
Systems Control, Inc.
1801 Page Mill Road
Palo Alto, California 94304

R. G. Alsmiller, Jr.
Oak Ridge National Laboratory
P. O. Box X
Oak Ridge, Tennessee 37830

Carl J. Anderson
Lawrence Livermore Laboratory
Box 808, MS L-216
Livermore, California 94550

David Andress
Science Management Corporation
8300 Professional Place
Landover, Maryland 20785

Anthony Apostolides
Jack Faucett Associates
5454 Wisconsin Ave., #1150
Chevy Chase, Maryland 20015

Robert Aster
Jet Propulsion Laboratory, MS 506-316
4800 Oak Grove
Pasadena, California 91103

M. Balasubramaniam
TRW, Inc.
8301 Greensboro Drive
McLean, Virginia 22102

Richard W. Barnes
Oak Ridge National Laboratory
P. O. Box X
Oak Ridge, Tennessee 37830

Jerry Bell
Washington Gas Light Co.
6801 Industrial Road
Springfield, Virginia 22151

Robert Bell
Lawrence Livermore Laboratory
Box 808, L-216
Livermore, California 94550

Yoav Benjamini
U. of Pennsylvania, Analysis Center
414-D Devereux Ave.
Princeton, New Jersey 08540

Javier Bernal
National Bureau of Standards
Center for Applied Mathematics
Washington, D. C. 20234

Robert L. Bivins
Los Alamos Scientific Laboratory
Mail Stop 605
Los Alamos, New Mexico 87545

K. O. Bowman
Union Carbide Corporation
Nuclear Division, 9704-1, Y-12
P. O. Box Y
Oak Ridge, Tennessee 37830

Tom S. Buckmeyer
Library of Congress/Congressional
Research Service
423 James Madison Building
Washington, D. C. 20540

Paul Bugg
Deptartment of the Interior
Office of Min. Policy & Res. Anal.
Washington, D. C. 20240

B. H. Burzlaff
Kentron International, Inc.
11317 Smallwood Drive
Burleson, Texas 76028

Charles Carpinella
Southern Connecticut Gas Co.
880 Broad Street
Bridgeport, Connecticut 06609

Robert T. Catlin
FERC, DOE, Room 7411
825 N. Capitol Street
Washington, D. C. 20426

Susan T. Cheng
U. of Pennsylvania, Analysis Center
3609 Locust Walk
Philadelphia, Pennsylvania 19104

Patrick D. Cheung
Atlantic Richfield Company
515 S. Flower Street
Los Angeles, California 90071

J. Phillip Childress
Department of Energy, Room 4324C
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

Harry Chmelynski
Jack Faucett Associates
5454 Wisconsin Ave., Suite 1150
Chevy Chase, Maryland 20015

Cecelia Chu
Department of Energy, Room 4324
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

Richard Collier
Systems Tech.
P. O. Box 459
Paonia, Colorado 81428

Dr. Burton H. Colvin
Dir., Center for Applied Mathematics
National Bureau of Standards
Washington, D. C . 20234

Thomas J. Cooney
3609 Locust Walk/C9
Wharton Analysis Center, U. of Penna.
Philadelphia, Pennsylvania 19104

Colleen M. Cornett
Department of Energy, Room 5317
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

Thomas A. Curran
Wharton Analysis Center
3609 Locust Walk
Philadelphia, Pennsylvania 19104

Ned Dearborn
DOE/EIA, M. S. 4530
NPO Bldg.
Washington, D. C. 20461

John A. Dearien
EG&G Idaho
Box 1625
Idaho Falls, Idaho 83401

Patricia R. Devine
Department of the Interior
1800 E Street, Room 4452
Washington, D. C. 20240

Paul Domich
National Bureau of Standards
Center for Applied Mathematics
Washington, D. C. 20234

Darryl J. Downing
Oak Ridge National Laboratory
P. O. Box X
Oak Ridge, Tennessee 37830

Anna M. S. Duncan
American Gas Association
1515 Wilson Blvd.
Arlington, Virginia 22209

Richard F. Dutcher
The Analysis Center
3609 Locust Walk
Philadelphia, Pennsylvania 19104

James A. Edmonds
Oak Ridge Associated Universities
11 DuPont Circle, N. W., Suite 805
Washington, D. C. 20036

Jane Ehrhardt-Johnson
University of Colorado
Boulder, Colorado 80309

Frank C. Emerson
EIA, Applied Analysis
Department of Energy, MS 4530
Washington, D. C. 20545

Robert T. Eynon
Department of Energy, Room 4530
1200 Pennsylvania Ave., N. W.
Washington, D. C. 20461

James Falk
National Bureau of Standards
Center for Applied Mathematics
Washington, D. C. 20234

Dr. Gene Finn
SEC
500 N. Capitol Street
Washington, D. C. 20549

J. S. Finucane
EIA
Room 7413 Federal Building
Washington, D. C. 20461

Dr. Eric C. Frankel
Science Management Corp.
8300 Professional Place
Landover, Maryland 20785

David Freedman
Statistics Department
University of California
Berkeley, California 94720

Neil Gamson
EIA, Department of Energy
Short-Term Division, Rm. 4324
Washington, D. C. 20461

Dr. Saul I. Gass
College of Business & Management
University of Maryland
College Park, Maryland 20742

William M. Gaynor
Pres., William M. Gaynor & Co., Inc.
10411 Tullymore Drive
Adelphi, Maryland 20783

Diana C. Gibbons
Wharton School Analysis Center
3609 Locust Walk, U. of Pennsylvania
Philadelphia, Pennsylvania 19104

Phyllis Gilmore
Department of Energy
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

Roger Glassey
EIA, U. of California
Dept. of Ind. Eng. & O. R.
Berkeley, California 94720

Alan J. Goldman
Johns Hopkins University
Mathematical Sciences Department
Baltimore, Maryland 21218

Jill Goldman
Wharton School Analysis Center
3609 Locust Walk, U. of Penna.
Philadelphia, Pennsylvania 19104

John W. Green
U. S. Dept. of Agriculture
Economics Dept., Colo. State Univ.
Ft. Collins, Colorado 80520

Rodney D. Green
Department of Energy
1811 Wyoming Ave., Room 28
Washington, D. C. 20009

Harvey Greenberg
Department of Energy, EIA
1200 Pennsylvania Ave., N. W.
Washington, D. C. 20461

Dr. Martin Greenberger
Mathematical Sciences Department
Johns Hopkins University
Baltimore, Maryland 21218

Rodney D. Green
Department of Energy
1811 Wyoming Ave., 21
Washington, D. C. 20009

Frank J. Gross
U. S. GAO
441 G Street, N. W., Room 5124
Washington, D. C.

James Gruhl
MIT, E38-436
77 Massachusetts Ave.
Cambridge, Massachusetts 02139

Hamid Habib-Agahi
JPL-Caltech, MS 506-316
4800 Oak Grove
Pasadena, California 91103

David Hack
Congressional Research Service
U. S. Library of Congress
Washington, D. C. 20540

Douglas R. Hale
EIA, Dept. of Energy, Rm. 7409
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

Malcolm R. Handte
Wharton Analysis Center
3609 Locust Walk
Philadelphia, Pennsylvania 19104

Carl M. Harris
Center for Management & Policy Res.
1625 I Street, N. W.
Washington, D. C. 20006

Jerry A. Hausman
MIT
E52-271A
Cambridge, Massachusetts 02139

John H. Herbert
OAOA/EIA/Applied Analysis
2929 Rosemary Lane
Falls Church, Virginia 22042

David S. Hirshfeld
David S. Hirshfeld Associates, Inc.
11057 Powder Horn Drive
Potomac, Maryland 20854

David Hochman
Wharton Analysis Center
3609 Locust Walk, U. of Penna.
Philadelphia, Pennsylvania 19104

Karla L. Hoffman
National Bureau of Standards
Center for Applied Mathematics
Washington, D. C. 20234

Kenneth C. Hoffman
Mathtech, Inc.
1611 N. Kent Street
Arlington, Virginia 22209

Peter Holihan
Department of Energy/CS
1H-031, Forrestal Bldg.
Washington, D. C.

Milton L. Holloway
Dir., Texas Energy Advisory Council
200 East 18th Street, Suite 513
Austin, Texas 78701

William A. Horn
3115 S. Hayes Street
Arlington, Virginia 22202

Paul F. Hultquist
U. of Colorado at Denver
1100 14th Street
Denver, Colorado 80202

Bruce G. Humphrey
Department of Commerce
Office of Regulatory Policy, Rm. 7614
Washington, D. C. 20230

Howard K. Hung
National Bureau of Standards
Center for Applied Mathematics
Washington, D. C. 20234

Richard H. F. Jackson
National Bureau of Standards
Center for Applied Mathematics
Washington, D. C. 20234

Lambert Joel
National Bureau of Standards
Center for Applied Mathematics
Washington, D. C. 20234

Charles R. Johnson
University of Maryland
Inst. for Physical Science & Tech.
College Park, Maryland 20742

Kent B. Joscelyn
HSRI Policy Analysis
University of Michigan
Ann Arbor, Michigan 48109

Byron G. Keep
Bonneville Power Administration
1500 NE Holliday Street
Portland, Oregon 97726

W. C. Kilgore
EIA
8506 Bromley Ct.
Annandale, Virginia 22003

Colleen Kirby
Analysis Center
3609 Locust Walk/C9
Philadelphia, Pennsylvania 19104

Andy S. Kydes
Brookhaven National Laboratory
NCAES, Bldg. 475
Upton, New York 11973

George M. Lady
Department of Energy
11307 Soward Drive
Kensington, Maryland 20795

Alexander H. Levis
MIT
35-410/LIDS
Cambridge, Massachusetts 02139

Andrew S. Loebl
Oak Ridge National Laboratory
P. O. Box X
Oak Ridge, Tennessee 37830

Maurice Long
The Orkand Corporation
8630 Fenton Street, Suite 938
Silver Spring, Maryland

Masood Malik
Appalachian Regional Commission
1666 Connecticut Ave., N. W.
Washington, D. C. 20235

Parvin Mansoure
Orkand Corp.
8630 Fenton Street
Silver Spring, Maryland 20910

William Marcuse
Brookhaven National Laboratory
Building 475
Upton, New York 11973

Martha Mason
MIT Energy Model Analysis Program
E38-570
Cambridge, Massachusetts 02138

John S. Maybee
University of Colorado
Dept. of Mathematics
Boulder, Colorado 80309

Dr. Lawrence S. Mayer
Analysis Center
Wharton School
Philadelphia, Pennsylvania 19104

Norman Miller
FERC
825 North Capitol Street
Washington, D. C.

Keith C. Mitchell
Science Management Corporation
15416 Durant St.
Silver Spring, Maryland 20904

Suzanne Mitchell
University of Pennsylvania
3609 Locust Walk/C9
Philadelphia, Pennsylvania 19104

Ellis A. Monash
U. S. Geological Survey
2080 Newcombe Drive
Lakewood, California 80215

Terry H. Morlan
Department of Energy, EIA
24309 Ridge Road
Damascus, Maryland 20750

Lincoln E. Moses
Department of Energy, EIA
Sequoia Hall
Stanford, California 94305

Frederic H. Murphy
Department of Energy, EIA
M. S. 4530
Washington, D. C. 20461

W. Charles Mylander
Department of Energy, EIA
Room 4530, Federal Bldg.
Washington, D. C. 20461

Nathaniel K. Ng
Department of Energy, Rm. 4517
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

David Nissen
Chase Manhattan Bank
One Chase Manhattan Plaza
New York, New York 10081

S. B. Noble
127 7th Street, S. E.
Washington, D. C. 20003

Robert Noorigian
TRW-Energy Systems
8301 Greensboro Drive
McLean, Virginia 22102

Richard P. O'Neill
Department of Energy, EIA
1200 Pennsylvania Ave., N. W.
Washington, D. C. 20461

Ramesh Parameshwaran
TRW Energy Systems Planning Division
8301 Greensboro Drive, #734
McLean, Virginia 22102

Shail Parikh
Oak Ridge National Laboratory
Integrative Energy-Economic Analysis
Energy Division
Oak Ridge, Tennessee 37830

Larry Parker
Congressional Research Service
Library of Congress, CRS/ENR
Washington, D. C. 20540

John D. Pearson
Department of Energy, EIA
2007 Franklin Ave.
McLean, Virginia 22101

David Pilati
Brookhaven National Laboratory
Bldg. 475
Upton, New York 11973

Clark Prichard
NRC, Mail Stop 1130 SS
Office of Research
Washington, D. C.

George W. Reinhart
NTIS
Springfield, Virginia 22161

David Reister
ORAV-IEA
P. O. Box 117
Oak Ridge, Tennessee 37830

Barbara C. Richardson
University of Michigan
Highway Safety Research Institute
Ann Arbor, Michigan 48109

Richard Richels
Electric Power Research Institute
3412 Hillview Ave.
Palo Alto, California 94303

Mark Rodekohr
Department of Energy, Room 4551
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

Paul F. Roth
Department of Energy, MS 4530
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

Lewis J. Rubin
EPRI, P. O. Box 10412
3412 Hillview Drive
Palo Alto, California 94303

Randi S. Ryterman
Wharton Analysis Center
3609 Locust Walk/C9
Philadelphia, Pennsylvania 19104

Prakash B. Sanghui
Jack Faucett Associates
5454 Wisconsin Avenue, Suite 1150
Chevy Chase, Maryland 20015

Rama Sastry
Office of Environment
Department of Energy
Washington, D. C.

Patsy Saunders
National Bureau of Standards
Center for Applied Mathematics
Washington, D. C. 20234

J. S. Seely
Department of Energy, EIA
1902 Bargo Ct.
McLean, Virginia 22101

Marquis R. Seidel
Department of Energy, FERC
825 N. Capitol Street
Washington, D. C. 20426

Roy Shanker
Hagler, Bailly & Co., Suite 350
2020 K Street, N. W.
Washington, D. C. 20006

Richard Shapiro
Kappa Systems
1409 Potter Drive
Colorado Springs, Colorado 80907

Susan H. Shaw
Department of Energy, M. S. 4530
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

Jeffrey E. Sohl
University of Maryland
10901 New Salem Ave.
Largo, Maryland 20870

A. Soyster
270 Whittemore
Virginia Tech, Department of IEOR
Blacksburg, Virginia 24061

Peter Stapleton
Southern Connecticut Gas
880 Broad Street
Bridgeport, Connecticut 06609

Martin M. Stein
ABT Associates
55 Wheeler Street
Cambridge, Massachusetts 02138

Leona Stewart
Oak Ridge National Laboratory
Bldg. G010, Rm. 203, P. O. Box X
Oak Ridge, Tennessee 37830

David Strom
Conservation Foundation
1717 Massachusetts Ave., N. W.
Washington, D. C. 20036

S. Scott Sutton
Department of Energy, EIA
M. S. 4530
Washington, D. C. 20461

James L. Sweeney
Stanford University
Energy Modeling Forum, Terman 406
Stanford, California 94025

T. Takayama
Department of Energy, EIA
Federal Building, MS 4530
Washington, D. C. 20641

Peter C. Tittmann
Analysis Center, Wharton School
U. of Pennsylvania, 3609 Locust Walk
Philadelphia, Pennsylvania 19104

John L. Trimble
Oak Ridge National Laboratory
P. O. Box X
Oak Ridge, Tennessee 37830

Leon L. Tucker
American Gas Association
1515 Wilson Blvd.
Arlington, Virginia 22209

Noel Uri
Department of Energy
OAOA/EIA
Washington, D. C. 20461

Christian Van Schayk
MVMA
300 New Center Bldg.
Detroit, Michigan 48202

Norman Wagoner
Dept. of Energy, FERC, Room 7102
825 N. Capitol Street, N. E.
Washington, D. C. 20426

Carol Wegrzynowicz
Department of Energy, EIA
12th & Pennsvylania Ave., N. W.
Washington, D. C. 20585

William I. Weinig
Department of Energy, EIA, Appl. Analysis
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

Charles Weisbin
ORNL, P. O. Box X
Bldg. B025, Room 20W
Oak Ridge, Tennessee 37830

Paul John Werbos
DOE, Room 7413, OEIV/EIA
Federal Bldg.
Washington, D. C. 20461

William B. Widhelm
University of Maryland
College of Business & Management
College Park, Maryland 20742

Bradford J. Wing
Department of Energy, MS 4530
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

Christoph Witzgall
National Bureau of Standards
Center for Applied Mathematics
Washington, D. C. 20234

David O. Wood
MIT
MIT Energy Lab.
Cambridge, Massachusetts 02139

Thomas J. Woods
General Accounting Office
Energy & Minerals Division, Rm. 5108
Washington, D. C. 20548

Douglas York
EEA, Inc.
1111 N. 19th Street
Arlington, Virginia 22204

Julie Zalkind
EIA, Applied Analysis, Rm. 4530
12th & Pennsylvania Ave., N. W.
Washington, D. C. 20461

NBS-114A (REV. 2-8C)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET (See instructions)	1. PUBLICATION OR REPORT NO. SP 616	2. Performing Organ. Report No.	3. Publication Date October 1981

4. TITLE AND SUBTITLE

Validation and Assessment of Energy Models
Proceedings of a Symposium held at the
National Bureau of Standards, Gaithersburg, MD, May 19-21, 1980

5. AUTHOR(S)

Saul I. Gass (Editor)

6. PERFORMING ORGANIZATION *(If joint or other than NBS, see instructions)*

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

7. Contract/Grant No.

8. Type of Report & Period Covered

Final

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS *(Street, City, State, ZIP)*

Energy Information Administration
Department of Energy
Washington, DC 20461

10. SUPPLEMENTARY NOTES

Library of Congress Catalog Card Number: 81-600087

☐ Document describes a computer program; SF-185, FIPS Software Summary, is attached.

11. ABSTRACT *(A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here)*

The Symposium on Validation and Assessment of Energy Models, held at the National Bureau of Standards (NBS), Gaithersburg, MD (May 19-21, 1980), was sponsored by the Energy Information Administration (EIA), of the Department of Energy (DOE), Washington, DC. The symposium was organized by NBS' Operations Research Division with a two-fold agenda: (1) to summarize the recent ideas and advances of model validation and assessment that have been applied to DOE energy models, and (2) to hold workshops on key open questions that are of concern to the validation and assessment research community. Speakers addressed current and future practices, the EIA model validation program, model structure and data, and model credibility. Full-day workshop sessions were held on the following topics: validating composite models, the measurement of model confidence, model structure and assessment, sensitivity and statistical analysis of models, and model assessment methodologies. This volume documents the symposium proceedings and includes the formal papers presented, discussant comments, panel discussions and questions and answers, and summaries of the issues and conclusions reached in the workshops.

12. KEY WORDS *(Six to twelve entries; alphabetical order; capitalize only proper names; and separate key words by semicolons)*

Assessment; composite models; data quality; energy models; mathematical models; model confidence; model credibility; policy models; sensitivity analysis; validation.

13. AVAILABILITY

[XX] Unlimited
☐ For Official Distribution. Do Not Release to NTIS
[XX] Order From Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.
☐ Order From National Technical Information Service (NTIS), Springfield, VA. 22161

14. NO. OF PRINTED PAGES

248

15. Price

$7.50

USCOMM-DC 6043-P80

NBS TECHNICAL PUBLICATIONS

PERIODICALS

JOURNAL OF RESEARCH—The Journal of Research of the National Bureau of Standards reports NBS research and development in those disciplines of the physical and engineering sciences in which the Bureau is active. These include physics, chemistry, engineering, mathematics, and computer sciences. Papers cover a broad range of subjects, with major emphasis on measurement methodology and the basic technology underlying standardization. Also included from time to time are survey articles on topics closely related to the Bureau's technical and scientific programs. As a special service to subscribers each issue contains complete citations to all recent Bureau publications in both NBS and non-NBS media. Issued six times a year. Annual subscription: domestic $13; foreign $16.25. Single copy, $3 domestic; $3.75 foreign.

NOTE: The Journal was formerly published in two sections: Section A "Physics and Chemistry" and Section B "Mathematical Sciences."

DIMENSIONS/NBS—This monthly magazine is published to inform scientists, engineers, business and industry leaders, teachers, students, and consumers of the latest advances in science and technology, with primary emphasis on work at NBS. The magazine highlights and reviews such issues as energy research, fire protection, building technology, metric conversion, pollution abatement, health and safety, and consumer product performance. In addition, it reports the results of Bureau programs in measurement standards and techniques, properties of matter and materials, engineering standards and services, instrumentation, and automatic data processing. Annual subscription: domestic $11; foreign $13.75.

NONPERIODICALS

Monographs—Major contributions to the technical literature on various subjects related to the Bureau's scientific and technical activities.

Handbooks—Recommended codes of engineering and industrial practice (including safety codes) developed in cooperation with interested industries, professional organizations, and regulatory bodies.

Special Publications—Include proceedings of conferences sponsored by NBS, NBS annual reports, and other special publications appropriate to this grouping such as wall charts, pocket cards, and bibliographies.

Applied Mathematics Series—Mathematical tables, manuals, and studies of special interest to physicists, engineers, chemists, biologists, mathematicians, computer programmers, and others engaged in scientific and technical work.

National Standard Reference Data Series—Provides quantitative data on the physical and chemical properties of materials, compiled from the world's literature and critically evaluated. Developed under a worldwide program coordinated by NBS under the authority of the National Standard Data Act (Public Law 90-396).

NOTE: The principal publication outlet for the foregoing data is the Journal of Physical and Chemical Reference Data (JPCRD) published quarterly for NBS by the American Chemical Society (ACS) and the American Institute of Physics (AIP). Subscriptions, reprints, and supplements available from ACS, 1155 Sixteenth St., NW, Washington, DC 20056.

Building Science Series—Disseminates technical information developed at the Bureau on building materials, components, systems, and whole structures. The series presents research results, test methods, and performance criteria related to the structural and environmental functions and the durability and safety characteristics of building elements and systems.

Technical Notes—Studies or reports which are complete in themselves but restrictive in their treatment of a subject. Analogous to monographs but not so comprehensive in scope or definitive in treatment of the subject area. Often serve as a vehicle for final reports of work performed at NBS under the sponsorship of other government agencies.

Voluntary Product Standards—Developed under procedures published by the Department of Commerce in Part 10, Title 15, of the Code of Federal Regulations. The standards establish nationally recognized requirements for products, and provide all concerned interests with a basis for common understanding of the characteristics of the products. NBS administers this program as a supplement to the activities of the private sector standardizing organizations.

Consumer Information Series—Practical information, based on NBS research and experience, covering areas of interest to the consumer. Easily understandable language and illustrations provide useful background knowledge for shopping in today's technological marketplace.

Order the **above** *NBS publications from: Superintendent of Documents, Government Printing Office, Washington, DC 20402.*

Order the **following** *NBS publications—FIPS and NBSIR's—from the National Technical Information Services, Springfield, VA 22161.*

Federal Information Processing Standards Publications (FIPS PUB)—Publications in this series collectively constitute the Federal Information Processing Standards Register. The Register serves as the official source of information in the Federal Government regarding standards issued by NBS pursuant to the Federal Property and Administrative Services Act of 1949 as amended, Public Law 89-306 (79 Stat. 1127), and as implemented by Executive Order 11717 (38 FR 12315, dated May 11, 1973) and Part 6 of Title 15 CFR (Code of Federal Regulations).

NBS Interagency Reports (NBSIR)—A special series of interim or final reports on work performed by NBS for outside sponsors (both government and non-government). In general, initial distribution is handled by the sponsor; public distribution is by the National Technical Information Services, Springfield, VA 22161, in paper copy or microfiche form.

U.S. DEPARTMENT OF COMMERCE
National Bureau of Standards
Washington, D.C. 20234

OFFICIAL BUSINESS

Penalty for Private Use, $300

POSTAGE AND FEES PAID
U.S. DEPARTMENT OF COMMERCE
COM-215

U.S.MAIL

SPECIAL FOURTH-CLASS RATE
BOOK